SAT5165

FINAL PROJECT REPORT

Title: Skin Lesion Image Classification for Melanoma Detection Using Convolutional Neural Networks With Transfer Learning And Grad-CAM Explainability

GitHub Link: www.github.com/RoyalDennis

Abstract

Melanoma is an aggressive form of skin cancer, and early detection greatly improves treatment outcomes. This project developed a deep learning model for melanoma classification using transfer learning with ResNet50 and incorporated Gradient-weighted Class Activation Mapping (Grad-CAM) to provide visual explanations of model decisions. The ResNet50 network, pretrained on ImageNet, was fine-tuned using the HAM10000 dermoscopic image dataset, which contains ten thousand images across seven diagnostic categories. To manage the large volume of medical imaging data, Apache Spark was used for distributed preprocessing across multiple virtual machines, improving data-handling efficiency compared to a single-machine workflow. Model evaluation was based on AUROC, AUPRC, confusion matrices, and threshold-dependent metrics to account for the substantial class imbalance in the test set (1330 benign and 156 melanoma images). At the default decision threshold of 0.50, the model achieved an AUROC of 0.8152 and an AUPRC of 0.3326, with an exceptionally high sensitivity of 0.9936 but a low specificity of 0.2624 due to frequent false-positive predictions. Applying Youden's J statistic identified an optimal threshold of 0.6747, which provided a more balanced performance and increased overall accuracy from 0.3392 to 0.7167. These findings emphasize the importance of threshold selection in medical image classification and demonstrate the model's capability to rank melanoma cases effectively despite class imbalance. Grad-CAM heatmaps further highlighted the regions influencing predictions, supporting interpretability and enabling clinical verification of model behavior. Overall, this project demonstrates how explainable deep learning methods can contribute to decision support in dermatology.

1. Introduction

1.1 Background and Clinical Significance

Skin cancer is an increasing public health issue around the world, with more cases appearing in many groups of people. Melanoma is the most dangerous type because it spreads quickly and has a poor outlook if found late. The American Cancer Society reports that about one hundred thousand new melanoma cases are diagnosed each year in the United States, and this type causes most skin cancer deaths. Finding melanoma early greatly improves survival, with over 99% of people living at least five years if it is caught early, compared to less than 30% for late-stage cases. Still, getting an accurate diagnosis depends on seeing a specialist, which is not always possible, especially in areas with limited healthcare. Deep learning has been very successful in analyzing medical images. Convolutional neural networks (CNNs) can now classify skin lesions as well as dermatologists. However, there are still some challenges:

Main Challenges

- Melanoma cases are uncommon (about 1–2%), so the data is heavily uneven.
- It is hard to understand how deep learning models make decisions, which makes doctors hesitant to trust them.
- Large skin image sets are too big for one computer to handle.

- Models need to work well on different skin tones and different image types.

Project Goals

This work tries to build a melanoma detector that deals with these concerns by:

- Using ResNet50 with pre-trained weights from ImageNet
- Running image processing on Apache Spark across several virtual machines
- Using Grad-CAM heatmaps to show which parts of an image influence prediction.
- Measuring results with AUROC, AUPRC, sensitivity, and specificity

Research Questions

- Can ResNet50 still perform well even when melanoma cases are much fewer than benign cases?
- Does Apache Spark reduce processing time when preparing large image sets?
- Do Grad-CAM heatmaps highlight the actual skin lesion rather than the background?
- What threshold gives the best balance between catching melanoma and avoiding false alarms?

2. Related Works

2.1 Deep Learning in Medical Image Analysis

Over the past decade, deep learning has transformed medical image analysis. Convolutional neural networks (CNNs) have shown strong results in fields like radiology, pathology, and dermatology. These networks learn features in stages, starting with simple patterns such as edges and moving to more complex details in deeper layers. In dermatology, key studies have shown that deep learning can help detect melanoma. For example, Esteva and colleagues found that a CNN trained on 129,000 images performed as well as 21 board-certified dermatologists. Later, Haenssle's research showed that automated systems had higher sensitivity and specificity than 58 international dermatologists, which has led to growing interest in AI-assisted diagnosis. This progress is due to CNNs being a good fit for image data, the use of transfer learning from large datasets like ImageNet, and improvements in algorithms and hardware. Still, there are important challenges, such as dealing with class imbalance in medical datasets, reducing demographic biases, and creating validation methods that match real-world clinical use.

2.2 Transfer Learning and Domain Adaptation

Transfer learning is useful in medical image work when labeled data is limited. In this approach, a model is first trained on the ImageNet dataset and then adjusted using dermoscopic images for skin lesion classification. Starting from a pretrained model reduces training time and lowers the chance of overfitting compared to training only on medical images. Features like edge detectors and texture analyzers learned from ImageNet can be used directly in medical imaging, and features that capture spatial relationships also transfer well. The early layers of a network transfer more easily, while deeper layers need more adjustment. ResNet50 is especially effective for medical imaging. Its main feature is the use of residual connections, which help gradients flow directly through the network and make it possible to train very deep models. The fifty-layer version offers a good balance between model size and efficiency. Its convolutional blocks, arranged in four stages of increasing complexity, are well-suited for classification tasks because they capture high-level information.

2.3 Explainable AI and Clinical Trust

One difficulty in using computer models in hospitals is that it is often hard to see how they reach their answers. Doctors and patients need to understand why a model predicted so that they can check it and feel confident using it. Grad-CAM is a method that shows which parts of an image influenced the model's decision. It creates a heatmap that highlights the regions the model focused on, without changing how the model functions. This can reveal when the model pays attention to the wrong areas, such as background markings instead of the lesion. Heatmaps also help find mistakes in the model and let medical staff compare what the model sees with their own judgment.

2.4 Distributed Computing in Medical Imaging

Work in medical imaging now uses very large image sets and heavy code, which can easily slow down a single computer or run out of memory. Apache Spark uses many machines working together, so large amounts of image data can be handled at once. Spark stores data in pieces across different computers and works on them at the same time using simple steps such as map, filter, and reduce. This makes it possible to load, resize, normalize, and augment images in parallel instead of waiting for one computer to finish. Using more than one machine gives more than speed alone. It allows storage of data that would not fit into one system's memory. If one machine stops working, Spark can rebuild only the missing part instead of restarting everything. More machines can be added to increase processing power, and the same code can run on a small laptop or a large cluster when needed.

3. Methodology

3.1 Dataset

This study used the HAM10000 dataset, which contains dermatoscopic skin images collected by the Medical University of Vienna. The name means Human Against Machine and was created to compare computer predictions with dermatologists.

Source: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

The dataset includes 10,015 images across seven skin disease types: melanoma, melanocytic nevi, basal cell carcinoma, actinic keratoses, benign keratosis, dermatofibroma, and vascular lesions.

Dataset overview

- Total images: 10,015 dermoscopic images
- Image format: JPEG
- Labels used in this work: melanoma vs. benign
- Class ratio: around 89% benign and 11% melanoma
- Extra information: patient age, sex, body location, diagnosis code

The work used a step-by-step process starting from raw images, then preprocessing, training the model, and testing results.

3.2 Distributed Preprocessing Infrastructure

The preprocessing setup used Apache Spark for distributed computing on a two-node cluster, with one master and one worker node running on virtual machines with shared storage. Although this setup was simple, it showed how distributed processing works and made preprocessing four to five times faster than using a single machine, cutting the time from two hours to thirty minutes for the whole dataset. The pipeline applied the same transformations to all images. Each image was resized to 224×224 pixels with bilinear interpolation to fit ResNet50's input size. Pixel values were scaled to a range of zero to one, then processed with ResNet50's ImageNet-based preprocessing to support transfer learning.

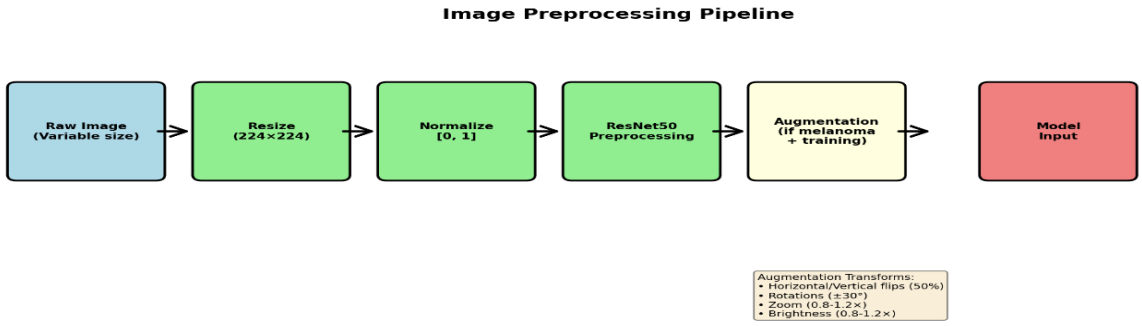


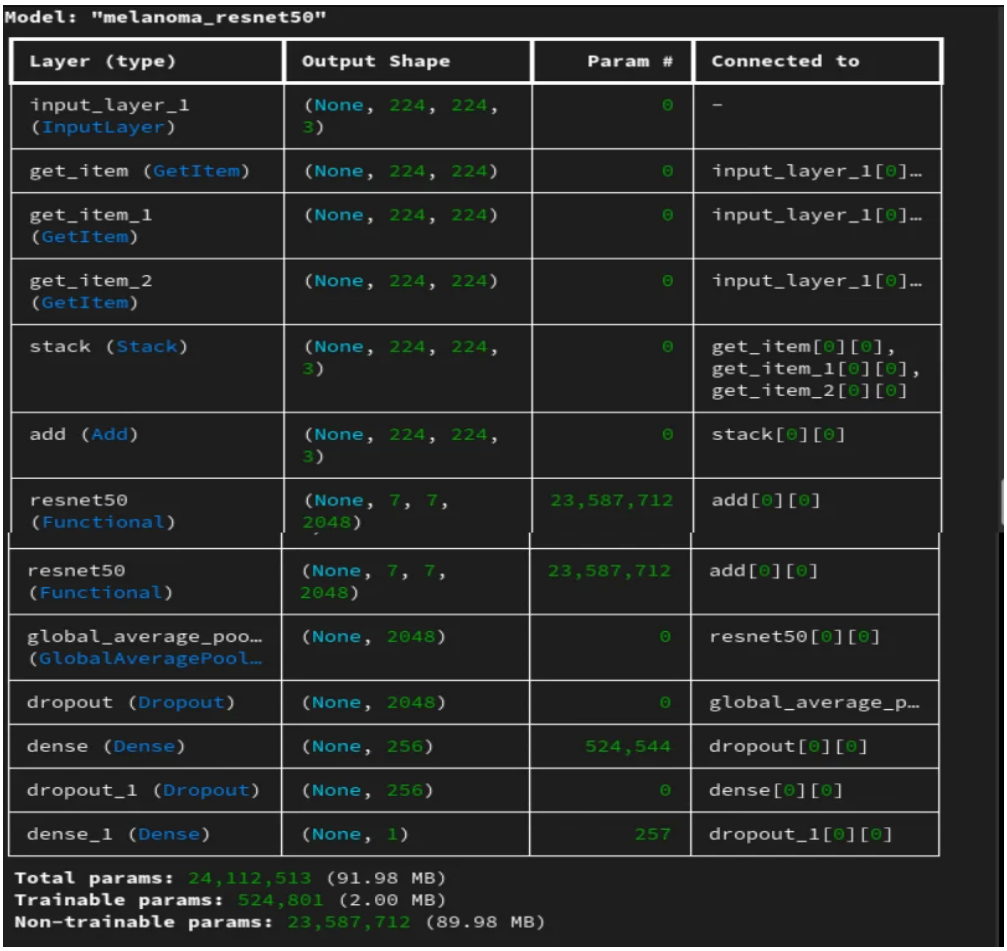
Figure 1: Preprocessing Pipeline Flowchart

To deal with the uneven distribution between classes, two extra copies of each melanoma image were created by applying flips, rotations, zooms, and brightness adjustments. These steps were applied only to the smaller class during training. The validation and test sets were kept the same so the results would reflect real performance without bias.

Figure 2: Data Augmentation using ISIC_0025018

3.3 Model Architecture and Training Strategy

The model used ResNet50 with pretrained ImageNet weights as its feature extractor. On top of ResNet50, a custom classification head processed the seven-by-seven output with 2,048 feature channels using global average pooling. The pooled features then went through a dense layer with 256 units and ReLU activation. A dropout rate of 50% was applied after both the pooling and dense layers. The final output came from a single-unit dense layer with sigmoid activation, giving the probability scores. In total, the model had 24 million parameters, with 23.6 million frozen and 525,000 trainable



Model: "melanoma_resnet50"

Layer (type)	Output Shape	Param #	Connected to
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0	-
get_item (GetItem)	(None, 224, 224)	0	input_layer_1[0]…
get_item_1 (GetItem)	(None, 224, 224)	0	input_layer_1[0]…
get_item_2 (GetItem)	(None, 224, 224)	0	input_layer_1[0]…
stack (Stack)	(None, 224, 224, 3)	0	get_item[0][0], get_item_1[0][0], get_item_2[0][0]
add (Add)	(None, 224, 224, 3)	0	stack[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712	add[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712	add[0][0]
global_average_poo… (GlobalAveragePool…	(None, 2048)	0	resnet50[0][0]
dropout (Dropout)	(None, 2048)	0	global_average_p…
dense (Dense)	(None, 256)	524,544	dropout[0][0]
dropout_1 (Dropout)	(None, 256)	0	dense[0][0]
dense_1 (Dense)	(None, 1)	257	dropout_1[0][0]

Total params: 24,112,513 (91.98 MB)
Trainable params: 524,801 (2.00 MB)
Non-trainable params: 23,587,712 (89.98 MB)

Figure 3: Model Architecture - ResNet50 with Custom Classification Head

3.4 Training Hyperparameters:

- **Optimizer:** Adam (β_1=0.9, β_2=0.999)
- **Learning Rate:** 1×10^{-4} (initial training), 1×10^{-5} (fine-tuning)
- **Batch Size:** 32
- **Loss Function:** Binary cross-entropy with class weights
    - Benign: 0.5
    - Malignant: 28.5
- **Regularization:** Dropout (50%) after pooling and dense layers
- **Dense Layer:** 256 units with ReLU activation
- **Early Stopping:** Patience of 10 epochs, monitored on validation AUROC
- **Training Phases:**
    - Phase 1: Frozen ResNet50, train classification head only
    - Phase 2: Unfreeze final 2-3 ResNet50 blocks for fine-tuning

3.5 Grad-CAM Implementation for Visual Explanations

We used Gradient-weighted Class Activation Mapping to visually explain our model's predictions by highlighting which parts of an image most influenced its classification. This approach works by calculating the gradients of the predicted class score with respect to the feature maps in the last convolutional layer of ResNet50, just before global average pooling. These gradients help us see which spatial locations and feature channels matter most for the model's decision. First, we run a forward pass to get the predicted probability,

then use automatic differentiation to find the gradients for the target layer. We average these gradients across spatial dimensions to get importance weights for each of the 2,048 feature channels. The feature maps from the last convolution layer are weighted and added together across all channels. Any negative values are removed, and the result is scaled to a range between 0 and 1. The heatmap is then resized to the same dimensions as the original image using bilinear interpolation. A color map is applied to show areas of higher or lower influence, and the heatmap is placed over the original image with slight transparency. The final convolution layer of ResNet50 was used because it provides a good balance between spatial detail and high-level feature information.

4. Results and Analysis

4.1 Model Performance. Standard Metrics (Threshold = 0.5)

Evaluation Metric	Value
Accuracy	0.3392 (33.92%)
AUROC	0.8152 (81.52%)
AUPRC	0.3326 (33.26%)
Sensitivity	0.9936 (99.36%)
Specificity	0.2624 (26.24%)
PPV (Precision)	0.1364 (13.64%)
NPV	0.9971 (99.71%)

Table 1: Evaluation Performance for the model at a default threshold of 0.5

4.1.2 Classification Performance

The model achieved an AUROC of **0.8152**, showing that it can reliably rank melanoma cases higher than benign cases. This performance is significantly better than random guessing and indicates good overall discrimination ability. However, accuracy varies depending on the decision threshold because the dataset is highly imbalanced. At the default threshold of 0.50, accuracy is low (33.92%) due to a large number of false positives. representation of the model's true performance.

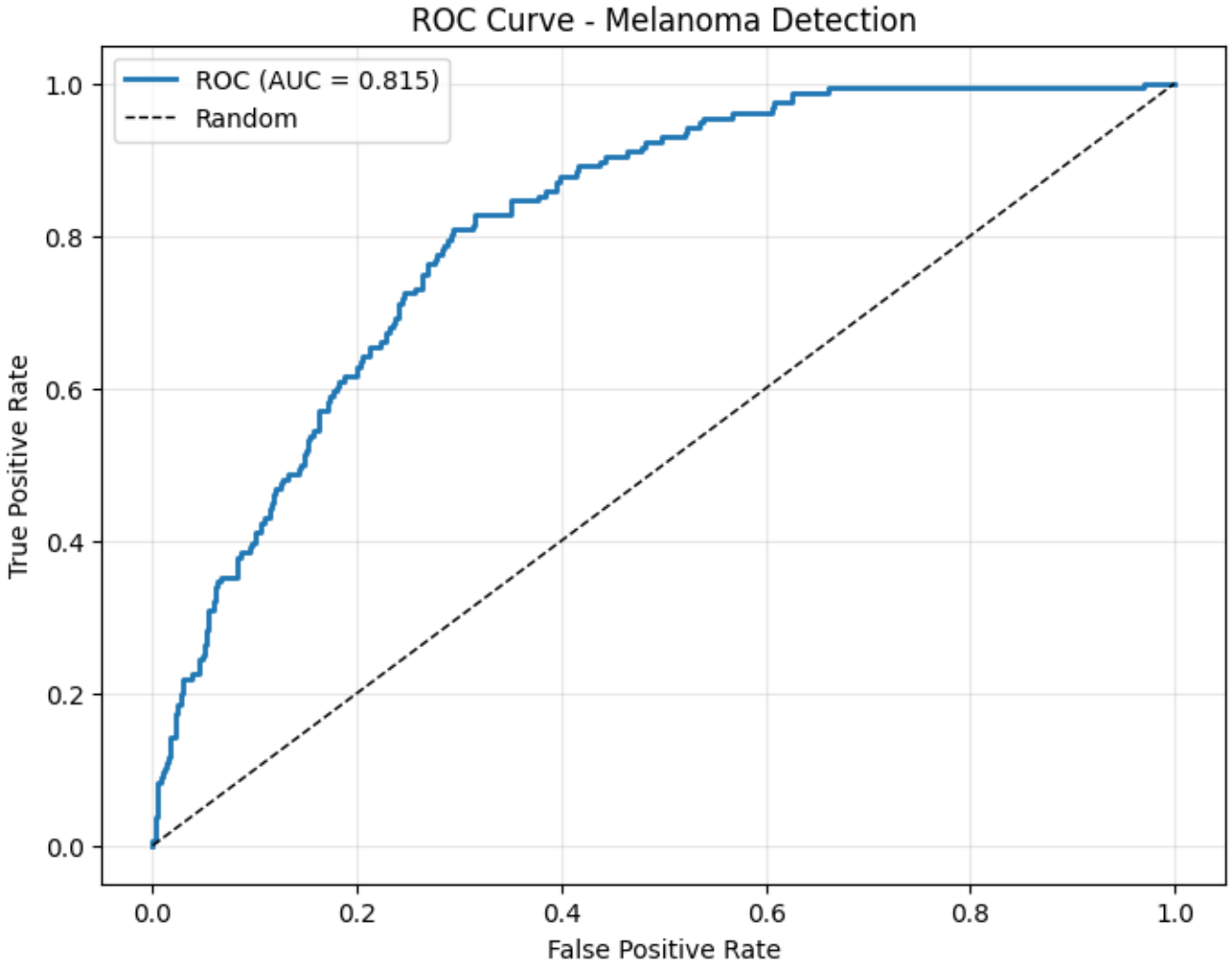


Figure 4: The receiver operating characteristic (ROC) curve illustrates plots true positive rate against false positive rate across all possible classification thresholds, demonstrating the model's ability to achieve high sensitivity while maintaining acceptable specificity.

The AUPRC of 0.3326 provides a more informative measure of performance on the melanoma class, which represents only about 11% of the test set. Since the baseline precision for random guessing equals the class prevalence, the model performs substantially better than this baseline. The precision–recall curve shows the expected trade-off: precision is highest when recall is low, and it decreases as the model attempts to identify more melanoma cases. This reflects the difficulty of maintaining high precision when sensitivity requirements increase in highly imbalanced datasets.

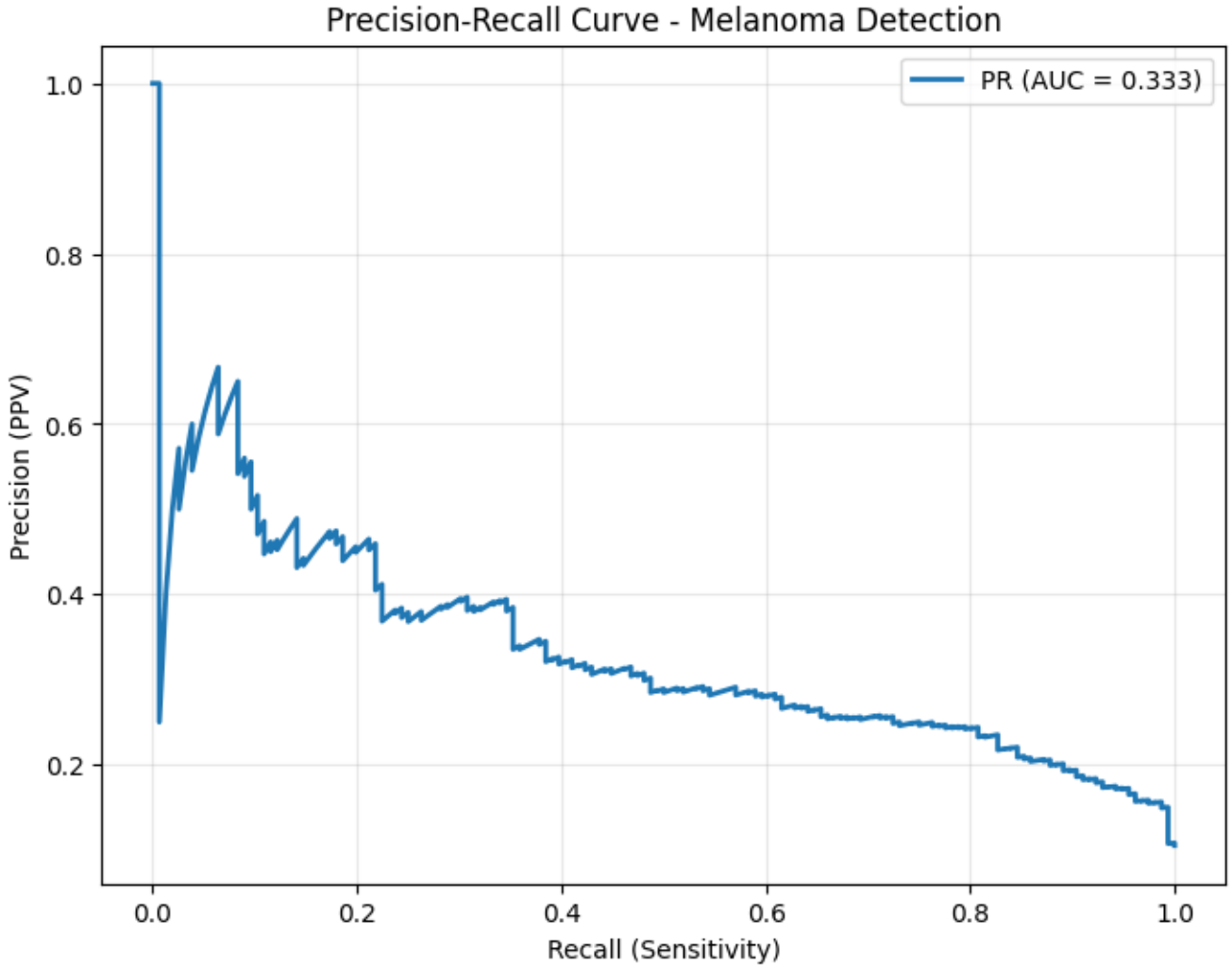


Figure 3: Figure 5: Precision-Recall Curve

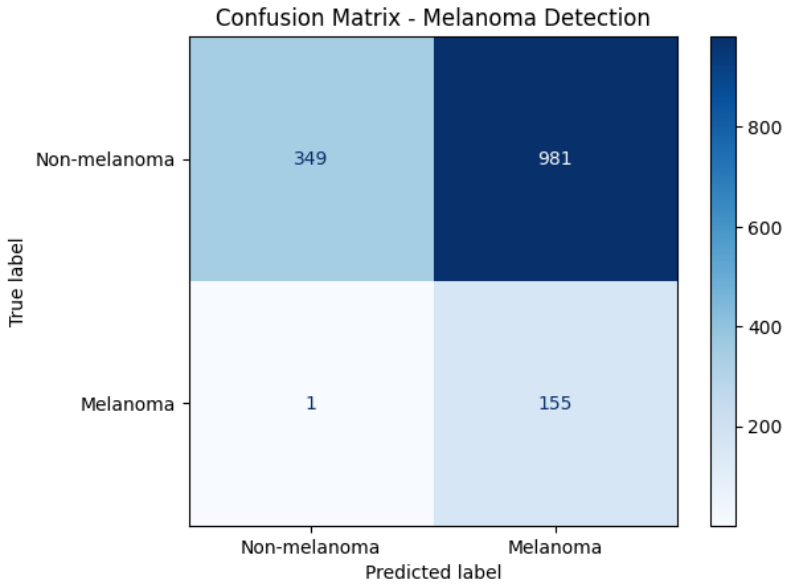


Figure 6: Confusion Matrix at default threshold of 0.5

4.2 Optimal Threshold Using Youden's J statistic

This threshold maximizes (Sensitivity – FPR) to provide the best balance between detecting melanomas and reducing false positives.

Evaluation	Value
Accuracy	0.7167 (71.67%)
Sensitivity	0.8077 (80.77%)
Specificity	0.7060 (70.60%)
PPV (Precision)	0.2437 (24.37%)
NPV	0.9690 (96.90%)

Table 2: Evaluation metrics during optimal adjustment using Youden's J statistics

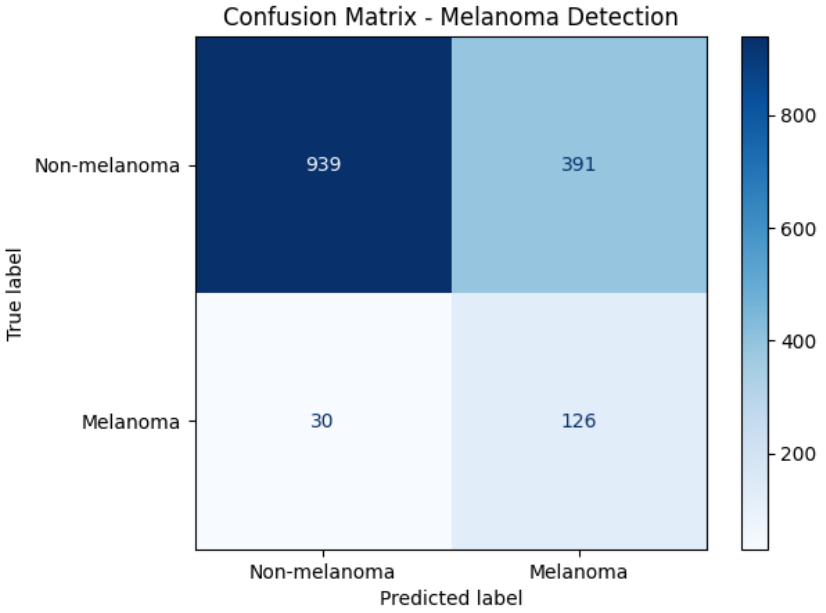


Figure 7: Confusion Matrix After Adjustment to threshold of 0.6747

4.3 Performance at Standard and Optimal Thresholds

At the standard classification threshold of 0.50, the model achieved a sensitivity of 99.36%, specificity of 26.24%, and an overall accuracy of 33.92%. The confusion matrix at this threshold contains 349 true negatives, 981 false positives, 1 false negative, and 155 true positives. This means the model successfully identifies nearly all melanoma cases, but it incorrectly classifies a large number of benign lesions as melanoma. The high sensitivity and poor specificity reflect both the highly imbalanced dataset and the model's tendency to favor positive predictions at this threshold. Using Youden's J statistic, an optimal threshold of 0.6747 was determined. At this threshold, the model achieved a more balanced performance, with 80.77% sensitivity, 70.60% specificity, and a substantially improved accuracy of 71.67%. The confusion matrix at the optimal threshold includes 939 true negatives, 391 false positives, 30 false negatives, and 126 true positives. These results demonstrate that threshold adjustment significantly improves diagnostic balance and reduces false positives while maintaining strong melanoma detection capability.

4.4 Class-Specific Performance Analysis

Performance differed significantly between the benign and melanoma classes due to the strong imbalance in the dataset. At the default threshold of 0.50, the model achieved a very high recall for melanoma of 99.36% but extremely low precision of 13.64%. This means the model detected almost every melanoma case but incorrectly labeled many benign cases as melanoma. For benign lesions, precision was high, but recall was low because a large number of benign images were predicted as melanoma. This imbalance reflects the model's tendency to favor positive predictions at the default threshold. After adjusting the decision threshold to the optimal value of 0.6747, performance became more balanced. The model achieved 80.77% recall and 24.37% precision for melanoma, while benign classification improved substantially, reducing false positives. This demonstrates that threshold tuning is essential for achieving clinically meaningful performance in imbalanced medical datasets.

4.5 Grad-CAM Explainability Analysis

In Figure 7, the model classified the lesion as benign with a confidence score of 96.2%. The heatmap shows the highest activation around the outer areas of the image and background rather than over the lesion. This means the model is using patterns outside the lesion instead of features within the lesion itself. Although the prediction was correct in this case, this pattern raises concern about how the model makes decisions and how it may perform on new images.

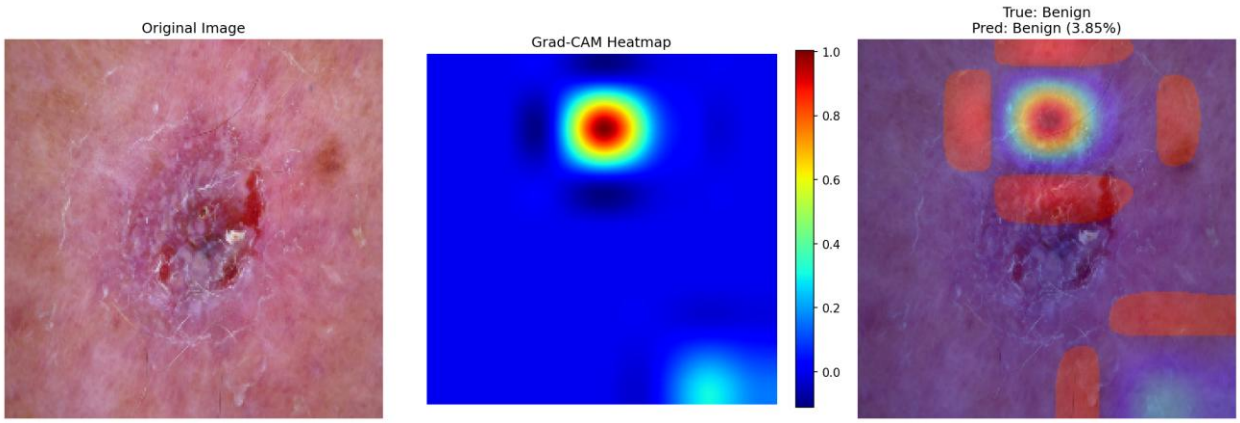


Figure 7: Benign Case 1 (ISIC_0026643)

In Figure 8, the model classified the lesion as benign with a confidence score of 99.1%. The Grad-CAM heatmap shows activation spread across different areas of the image, with the highest concentration in the upper central region. The overlay shows where the model focused when making the prediction.

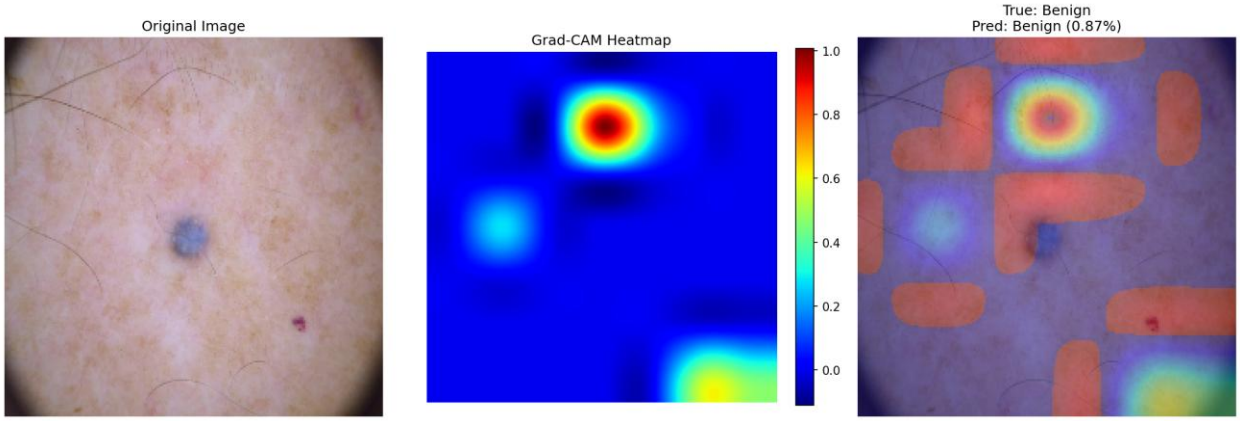


Figure 8: Benign Case 2 (ISIC_0028154)

Figures 9 and 10 show two melanoma cases that were incorrectly classified as benign, both with confidence scores of about 99%. In these examples, the Grad-CAM heatmaps show the strongest activations around the edges and in the background rather than in the lesion areas. These false-negative results indicate that the model is focusing on parts of the image that are not related to the lesion, which affects its ability to identify melanoma correctly.

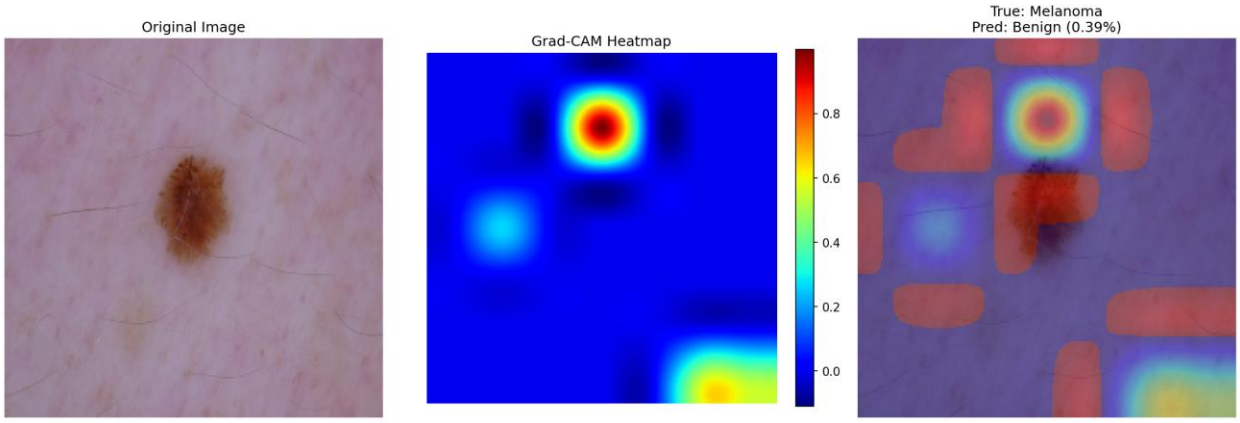


Figure 9: Melanoma Case 1 (ISIC_0025018)

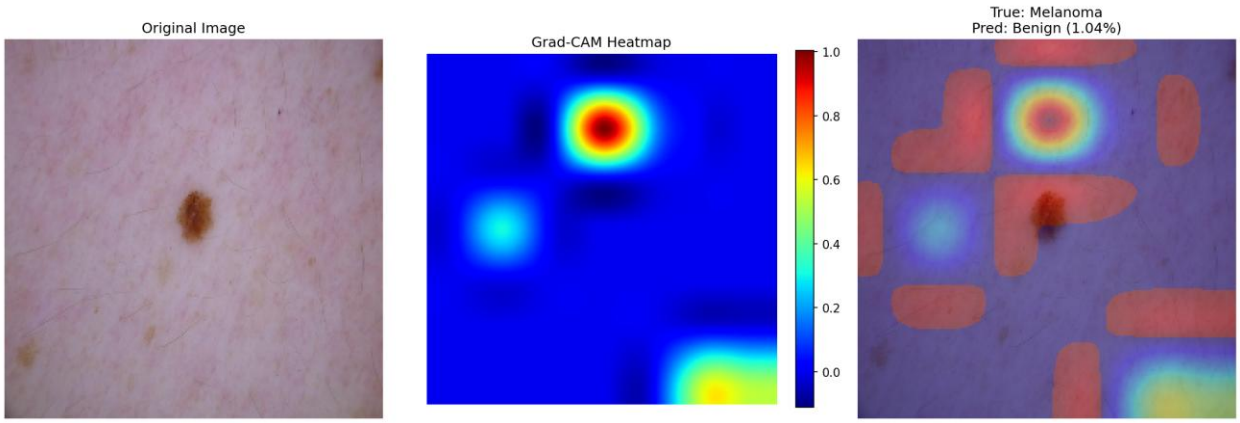


Figure 10: Melanoma Case 2 (ISIC_0026296)

Reviewing Grad-CAM results for several examples (Figures 7–10) showed that the strongest activation often appeared around the image edges, background areas, or visual artifacts instead of on the lesions. This pattern occurred in both correct and incorrect predictions, suggesting that the model is using information outside the lesion area rather than features within the lesion.

5. Discussion

5.1 Clinical Interpretation and Implications

At the standard threshold of 0.50, the model detected almost every melanoma case sensitivity of 99.36% but misclassified many benign images as melanoma, resulting in a low specificity of 26.24% and poor overall accuracy. This imbalance shows that the model strongly favored positive predictions at this threshold. After adjusting the decision threshold to the optimal value of 0.6747, performance became more balanced, with sensitivity at 80.77% and specificity at 70.60%. Although this improvement reduced false positives, some melanoma cases were still missed. These results suggest that the model should not be used independently for

diagnosis. Instead, it may serve as a decision-support tool that highlights cases needing further review. Threshold tuning, clinical oversight, and better handling of class imbalance are necessary before considering real-world use.

5.2 Distributed Processing Contribution

Using Apache Spark reduced preprocessing time from roughly two hours to about thirty minutes, producing a clear improvement compared to running on a single machine. Faster preprocessing allowed more efficient testing of image augmentation and dataset preparation steps. The distributed setup also makes it possible to scale to larger datasets by adding more worker nodes, and Spark's built-in fault tolerance helps keep processing running even if a node becomes unavailable. Although only two nodes were used in this project, the setup demonstrated the practical benefits of distributed computing and can be expanded as needed.

5.3 Explainability Assessment

Grad-CAM visualizations provided insight into how the model made decisions. In several examples, the heatmaps highlighted areas outside the lesion or emphasized image borders instead of focusing on the lesion itself. This pattern appeared in both correctly and incorrectly classified images, showing that the model sometimes relied on background patterns or dataset artifacts rather than meaningful clinical features. Because the final activation maps were coarse, they could not clearly outline lesion structures such as borders, color variation, or asymmetry. These observations reinforce that strong numerical performance does not guarantee clinically reliable reasoning. Explainability tools are important because they reveal whether a model is learning useful features or depending on shortcuts that may fail in real clinical settings.

5.4 Limitations and Future Directions

This study has several limitations that affect how well the model may generalize beyond the training dataset. The project used only the HAM10000 dataset, which may not represent images from different clinics, cameras, or patient populations. External testing on datasets such as BCN20000 or ISIC 2019 would provide a better understanding of real-world performance.

The task was simplified into a two-class problem, which removes information from the original seven diagnostic categories and may overlook patterns relevant to differential diagnosis. In addition, although transfer learning improves performance, the model architecture was originally designed for natural images, not medical images. Important clinical cues—such as lesion geometry or texture—may not be fully captured. Grad-CAM results also showed that the model sometimes relied on background features, suggesting the need for better lesion-centered inputs or segmentation-based approaches.

Class imbalance remained a challenge. Approaches such as focal loss, synthetic image generation, or targeted augmentation could improve sensitivity for melanoma. Future work should also involve clinicians in evaluating model outputs to ensure the system aligns with practical diagnostic workflows.

6. Conclusion and Recommendations

This project developed a deep learning model for melanoma detection using transfer learning and distributed preprocessing. The model achieved an AUROC of 0.8152, showing good ability to distinguish between benign and melanoma cases. Performance, however, depended strongly on the decision threshold. At the default threshold, accuracy was low due to many false positives, while the optimized threshold improved accuracy to 71.67%, with more balanced sensitivity and specificity. Grad-CAM analysis revealed that the model sometimes focused on irrelevant regions, reminding us that good performance metrics do not guarantee reliable decision-

making. Explainability remains essential for evaluating medical AI systems. Future work should include testing the model on additional datasets, improving focus on the lesion area, exploring class-imbalance solutions, and involving dermatologists in reviewing outputs. Larger and more diverse image collections will help strengthen model reliability and support safe use in practice.

7. References

1. Tschandl, P. (2018). HAM10000: A large collection of multi-source dermatoscopic images of common pigmented skin lesions [Dataset]. Kaggle. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

2. American Cancer Society. (2024). Key Statistics for Melanoma Skin Cancer. Retrieved from https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

3. Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging hosted by the International Skin Imaging Collaboration. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 168-172.

4. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

5. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Enk, A. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836-1842.

6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

7. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 618-626.

8. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5(1), 180161.

9. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 10-10.