Skin Cancer Detection Using Convolutional Neural Networks

Pedro Silva (n.º 64926)
Pedro Diz (n.º 64852)
Deep Learning, FCUL

Abstract

For the milestone 2 of the project, we present an upgrade over the minimal deep-learning pipeline for binary melanoma detection developed in milestone 1 relative to the ISIC 2018 Task 3 dataset of 11 720 dermoscopic images. After patient-wise splitting (70/15/15), images are cropped, resized, normalized, and augmented before fine-tuning an EfficientNet-B0. With these upgrades and experiments, we can confidently conclude that overall performance improved, the AUROC rose by 2–3 points over the previous baseline.

Keywords: deep learning, dermoscopy, melanoma, skin cancer, EfficientNet, focal loss

1. Introduction

Portugal has a hot climate and extensive beach culture, leading to high sun exposure, which significantly increases the risk of skin cancer[1]. Early detection is crucial, as it greatly improves the chances of successful treatment while reducing long-term damage and increasing survival rates[2]. We propose a solution for detecting cancerous cells using a convolutional neural network. Our approach relies exclusively on the ISIC 2018 Archive, as the originally considered HAM10000 dataset was found to contain duplicate images from ISIC. The pipeline includes data acquisition through the ISIC API, preprocessing, model training, and performance evaluation. The source code is present at *https://github.com/PedroDiz/AP-2025*

2. Problem Statement

We address the task of binary classification of dermoscopy skin-lesion images into benign (0) versus malignant (1). Our dataset consists of 11 720 RGB images of size 600×450 pixels drawn from the ISIC 2018 Task 3 challenge.We ignored the challenge's original splits and instead perform a patient-wise stratified split into 70 % training, 15 % validation, and 15 % test. The resulting class distribution is highly imbalanced (figure 5): 81.37 % benign and 18.63 % malignant.

At inference time, the model takes a 600×450 image as input and outputs a malignancy probability $p \in [0, 1]$. Initially, we naively binarize this probability at 0.5, and later evaluate using the Youden point to find the optimal threshold.

We evaluate performance **primarily** using the **AUROC**, which is the standard benchmark metric in dermoscopy and is widely used in related research. **Secondarily**, we report **balanced accuracy** to account for class imbalance. Additional evaluation strategies include the **F$_1$-score** and **recall**.

3. Technical Approach

3.1. Data Pipeline

We downloaded all images from the ISIC 2018 Task 3 API, split them into multiple ZIP files, and stored them in our repository. In future runs, these ZIP files are downloaded and extracted into a single folder. We then create a single metadata file (`lesions.csv`) containing three columns: image filename, patient ID, and binary label (0=benign, 1=malignant). We performed a patient-wise split into 70 % training, 15 % validation, and 15 % test sets.

Prior to training, each image undergoes the following preprocessing steps:

1. Center-crop to square and resize to 224×224 pixels.
2. Convert to tensor and normalize per-image to zero mean and unit variance.
3. On-the-fly data augmentations *during training only*: random resized crop (scale 0.9–1.0), horizontal and vertical flips, rotation $\pm 15°$, color jitter (brightness/contrast/saturation 0.1), and random zoom.

We acknowledge that employing a 5-fold GroupKFold cross-validation at the patient level would have ensured that each patient appears exactly once in the test set across all folds, yielding:

[1] https://www.ipolisboa.minsaude.pt/noticias/cancro-da-pele-e-dos-mais-frequentes-em-portugal/

[2] https://www.cas.org/resources/cas-insights/how-biomarkers-unlock-faster-cancerdetection-improving

- Reduced risk of data leakage between training and evaluation splits,
- More reliable estimates of model generalisation performance,
- More stable variance estimates across folds.

However, due to time constraints and limited GPU resources, we were unable to implement this approach in a timely manner. Consequently, we retained the original 70%/15%/15% (train/validation/test) split pipeline established in Milestone 1.

3.2. Model Architecture and Training Setup

For the second baseline, we used a CNN backbone from the `timm` library, **EfficientNet-B0**, as we did on the milestone 1, pretrained on ImageNet. The raw model outputs logits z, which are converted to probabilities $p = \sigma(z)$ via a sigmoid activation. Excusively for this baseline, we added:

- **Mixed-precision + grad-acc:** halves memory usage and doubles effective batch size.
- **Balanced sampler(60/40):** equalizes class exposure each batch, improving minority-class learning.
- **Warm-up freeze:** trains only the new head initially for stable fine-tuning of the pretrained backbone.

All other settings remain as in Milestone 1 (which used a highly imbalanced dataset), but we've reduced the focal BCE *alpha* from 0.75 to 0.6 because we now apply weighted random sampling to up-sample the minority class.

- **Loss function**: We use Focal Binary Cross-Entropy, that up-weights the malignant class relative to the benign class. It extends standard BCE by:
  - α: a class-weight scalar (set to 0.6)
  - γ: the focusing parameter (set to 2)
- **Optimizer**: AdamW with learning rate 3×10^{-4} and weight decay 1×10^{-4}.
- **Scheduler**: Cosine annealing over 30 total epochs.
- **Batch size**: 32.
- **Device**: GPU (NVIDIA T4).

3.3. Inference and Metrics

During validation and testing, we compute the raw logits z, convert to probabilities $p = \sigma(z)$, and binarize at threshold $t = 0.5$. We record:

- **AUROC**: area under the full ROC curve.
- **Balanced accuracy**: average of sensitivity and specificity.
- **Recall**: The ratio between the number of Positive samples correctly classified as Positive to the total number of Positive samples.
- **F$_1$-score**: harmonic mean of precision and recall on the malignant class.
- **Confusion matrix**: to visualize true/false positives and negatives.

A schematic of the pipeline is shown in Figure 3.

4. Evaluation Milestone 2

4.1. Analysis of Baseline Performance

Loss Curve Analysis. As the previous baseline of phase 1, both the training and validation losses decrease steadily and remain closely aligned throughout all 20 epochs (figure 1), with no widening gap indicative of overfitting. Moreover, there is no clear trend toward a plateau in the validation loss, suggesting that the model has not yet reached its capacity for improvement. Consequently, our current run is to short both to identify an optimal stopping point and to observe diminishing returns. Note that all threshold-dependent metrics were calculated using Youden's point.

Ranking Metrics.

- **AUROC**: 0.909, up from 0.88 at baseline.
- **Macro-F$_1$ Score**: 0.778, exceeding the 0.72 baseline.
- **Balanced Accuracy**: 0.755, improving on the 0.71 baseline.
- **Recall**: 0.56 (not calculated in Phase 1 of the project).
- **Confusion Matrix** confirms the results obtained from the other metrics (Figure 2).

Grad-CAM visualizations (Figure **??**) were used to qualitatively validate the model's predictions

Conclusion. We conclude that the new baseline outperforms the previous one: its improved metrics likely result from the weighted sampling used to correct the dataset's pronounced benign-class imbalance.

4.2. Optimal Threshold Selection

To enhance classification performance beyond the default 0.5 cutoff, we computed Youden's J statistic,

$$J(t) = \mathrm{TPR}(t) - \mathrm{FPR}(t),$$

and used it to select the optimal decision threshold for our metrics, resulting in improved overall performance.

4.3. Experiments

4.3.1. Baseline with Image Resizing to 300x300 15 epochs

This experiment yielded an AUROC of **0.906**, balanced accuracy of **0.727**, F1-score of **0.753**, and recall of **0.502**. Despite the reasonable metrics, this approach underperformed compared to the baseline. We believe the model may have required more training time to effectively learn the patterns, suggesting that increasing the number of epochs could improve performance.

4.3.2. Baseline with Mixup 0.2 (15 Epochs)

This experiment aimed to address class imbalance by applying mixup with a factor of 0.2. It achieved an AUROC of **0.884**, balanced accuracy of **0.674**, F1-score of **0.701**, and recall of **0.397**. Performance was lower than the baseline. This may be attributed to the limited training duration (15 epochs) or the use of a weaker model architecture compared to the B3 variant.

4.3.3. Baseline with Shades of Grey Preprocessing (15 Epochs)

This experiment introduced grayscale transformations to enhance model robustness by reducing reliance on color features. It produced an AUROC of **0.874**, balanced accuracy of **0.697**, F1-score of **0.725**, and recall of **0.443**. The results were lower than the baseline, indicating that this transformation did not improve performance in this case. This may be due to the reduced color information making the task more difficult, requiring a stronger model or longer training to compensate.

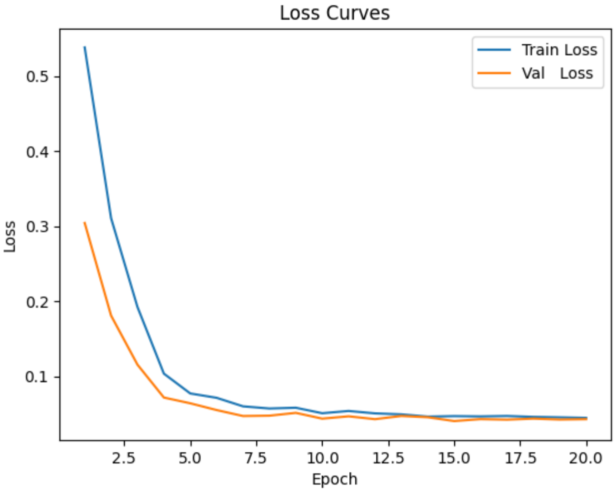

Figure 1. Training and validation loss curves over 10 epochs.

5. Future Work

In future work, once this baseline is solid, we will explore:

- Utilizing 5-fold patient GroupKFold.
- Implement facebook/dinov2-vit-s-14 experiment.

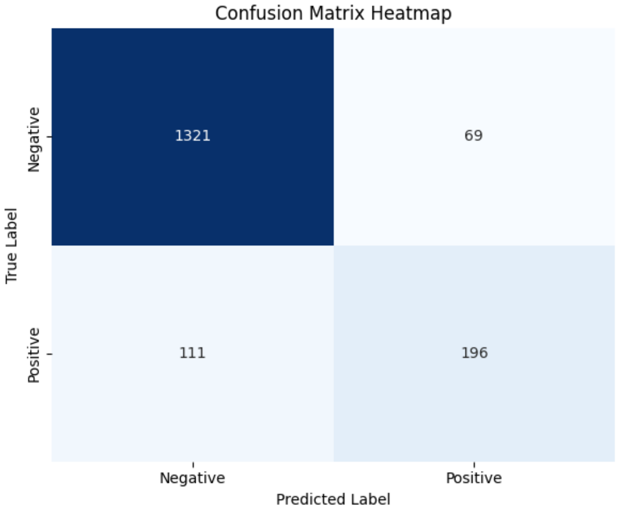

Figure 2. Confusion Matrix

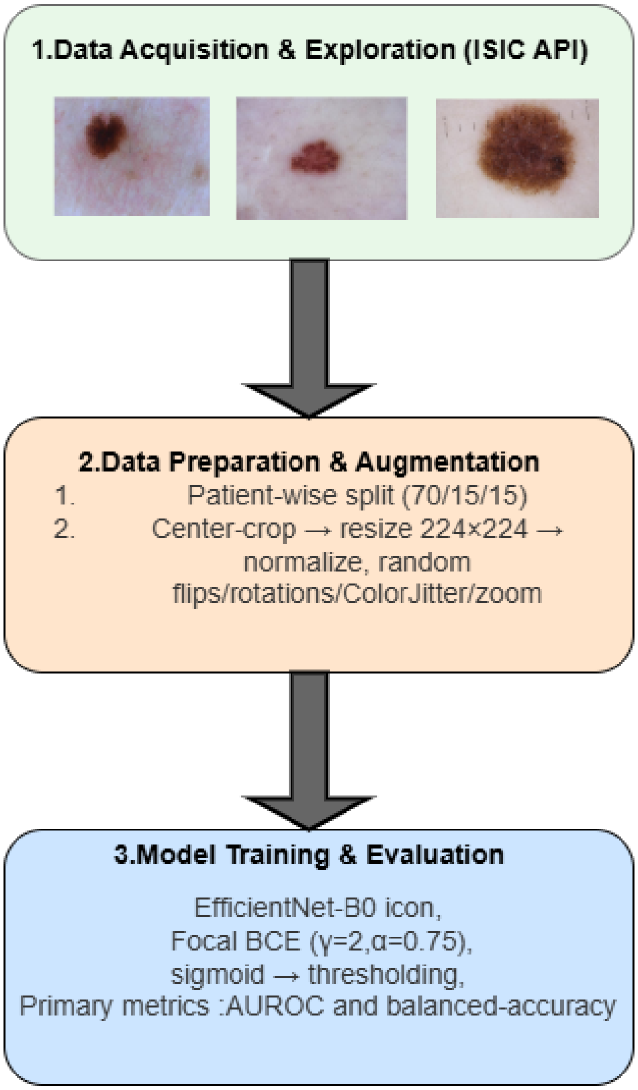
1.Data Acquisition & Exploration (ISIC API)
2.Data Preparation & Augmentation
1. Patient-wise split (70/15/15)
2. Center-crop → resize 224×224 → normalize, random flips/rotations/ColorJitter/zoom
3.Model Training & Evaluation
EfficientNet-B0 icon,
Focal BCE (γ=2,α=0.75),
sigmoid → thresholding,
Primary metrics :AUROC and balanced-accuracy

Figure 3. Pipeline

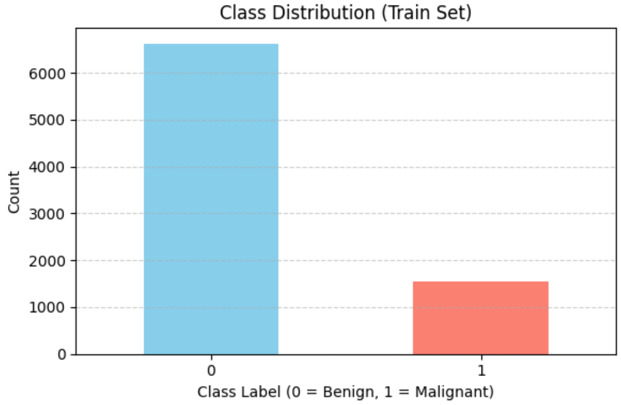

Figure 4. Dataset Histogram

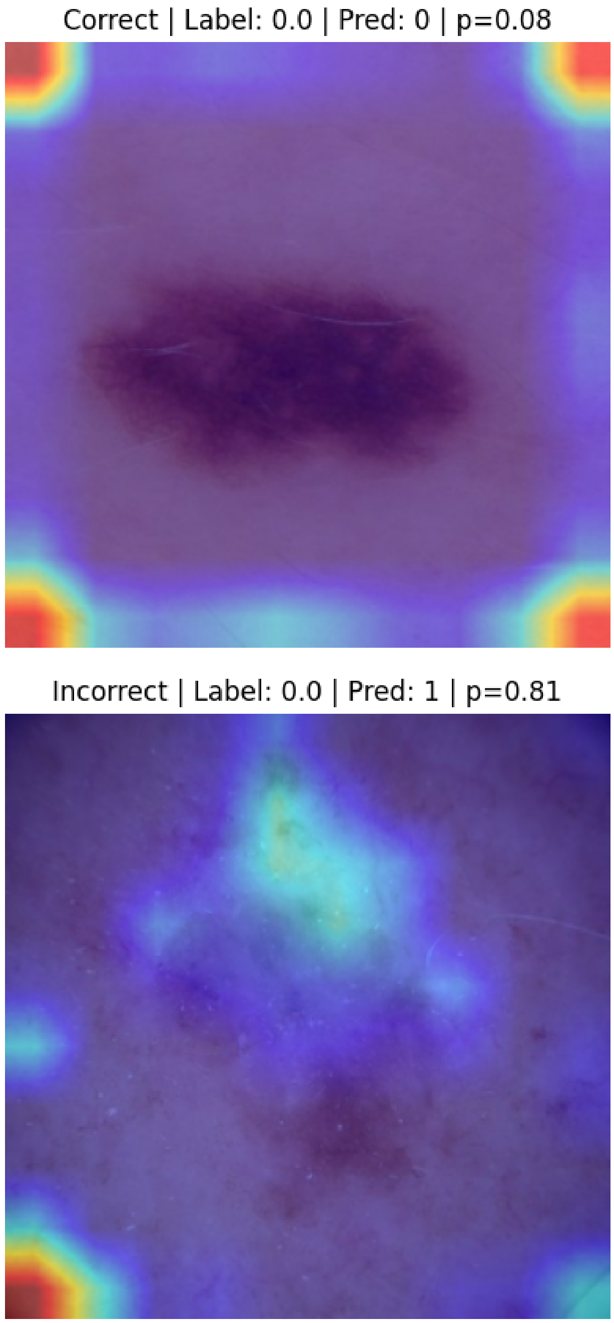
Correct | Label: 0.0 | Pred: 0 | p=0.08
Incorrect | Label: 0.0 | Pred: 1 | p=0.81

Figure 5. Top: Correct prediction. Bottom: Incorrect prediction