

Artificial Intelligence in Mathematical Engineering

PROJECT REPORT

AI-Assisted Diagnosis on HAM10000

The Effect of Loss Functions and Sampling on
Imbalanced Skin-Cancer Classification

Author: Ömer Kahyaoğlu

E-mail: kahyaoglu20@itu.edu.tr

Date: May 5, 2026

Student Number: 090240741

Mathematical Engineering Undergraduate Programme
Istanbul Technical University

MYZ 309E CRN: 23577

Class Lecturer: Professor Muhammed Kurulay, PhD

Problem Statement

On the HAM10000 dermoscopic dataset (10 015 images, 7 diagnostic classes), the majority class *melanocytic nevi* (**nv**) accounts for $\sim 67\%$ of the data while the rarest class *dermatofibroma* (**df**) accounts for only $\sim 1.1\%$. A model trained naively with Cross-Entropy will optimise mean log-likelihood and trivially achieve high *accuracy* by mostly predicting **nv** — while missing the rare and clinically important classes. We compare **five** loss/sampling configurations under an identical CNN, optimiser and seed on the full real HAM10000 corpus, and report *macro-F1* and *per-class recall* (the metrics that an imbalanced clinical task actually needs). Headline result: **Focal loss ($\gamma=2$) reaches macro F1 = 0.443 vs. 0.381 for Standard CE**, and is the only loss-side fix that recovers non-zero recall on the rarest class **df**.

1. Introduction

The HAM10000 collection [1] contains 10 015 dermoscopic images of pigmented skin lesions, hand-labelled across seven diagnostic classes: melanocytic nevi (**nv**), melanoma (**mel**), benign keratosis-like lesions (**bk1**), basal cell carcinoma (**bcc**), actinic keratoses (**akiec**), vascular lesions (**vasc**) and dermatofibroma (**df**). Figure 1 shows that the dataset is *severely imbalanced*: a constant predictor that always outputs **nv** would already attain $\approx 67\%$ accuracy — which makes plain accuracy a misleading optimisation target.

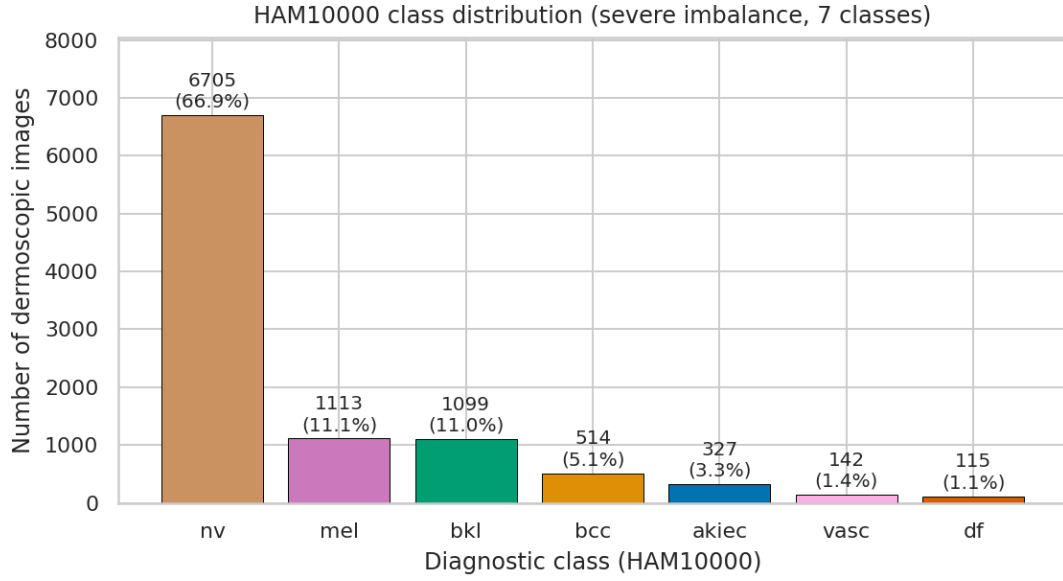


Figure 1: Class distribution of the HAM10000 training data. The **nv** class dominates ($\approx 67\%$), while **df** accounts for only $\approx 1.1\%$ of the corpus.

The objective of this project is therefore not just to *classify* skin lesions, but to *study how the choice of loss function and sampling strategy interact with class imbalance*: do techniques that explicitly re-weight rare classes recover their recall, and at what cost to majority-class performance?

2. Methodology

2.1. Pipeline overview and CNN architecture

The full pipeline (Figure 2) is: (i) load all 10 015 HAM10000 JPGs from the official Harvard Data-verse zips; (ii) stratified 70/15/15 train/validation/test split (per-class proportions preserved); (iii) 64×64 resize, ImageNet-style mean/std normalisation and light training-time augmentation (horizontal/vertical flip); (iv) training of a from-scratch CNN ($\sim 106k$ parameters) for 8 epochs with Adam; (v) test-set evaluation with imbalance-aware metrics.

Input ($3 \times 64 \times 64$) $\rightarrow$ **Conv** $3 \rightarrow 16$ + BN + ReLU + MaxPool₂ ($16 \times 32 \times 32$) $\rightarrow$ **Conv** $16 \rightarrow 32$ + BN + ReLU + MaxPool₂ ($32 \times 16 \times 16$) $\rightarrow$ **Conv** $32 \rightarrow 64$ + BN + ReLU + MaxPool₂ ($64 \times 8 \times 8$) $\rightarrow$ **Conv** $64 \rightarrow 128$ + BN + ReLU + MaxPool₂ ($128 \times 4 \times 4$) $\rightarrow$ **GlobalAvgPool** (128) $\rightarrow$ **Dropout**_{0.4} $\rightarrow$ **FC** $128 \rightarrow 64$ + ReLU $\rightarrow$ **Dropout**_{0.4} $\rightarrow$ **FC** $64 \rightarrow 7$ (logits)

Figure 2: CNN architecture. Four Conv / BatchNorm / ReLU / MaxPool blocks followed by a Global Average Pooling head and a small FC head. Total: **106 391** trainable parameters.

2.2. Why this loss family, this optimiser, these hyperparameters?

Key Theory & Formulas

With class counts $\{n_c\}_{c=1}^C$ and per-sample softmax probability $p_t(x_i) = \hat{p}(y_i | x_i)$, the four imbalance-handling losses studied here are

$$\mathcal{L}_{\text{CE}}(x, y) = -\log p_t, \quad \mathcal{L}_{\text{wCE}}(x, y) = -\frac{1/n_y}{\sum_c 1/n_c} C \log p_t, \quad (1)$$

$$\mathcal{L}_{\text{focal}}(x, y) = -\alpha_y (1 - p_t)^\gamma \log p_t \quad (\gamma = 2), \quad \mathcal{L}_{\text{CB}}(x, y) = -\frac{1 - \beta}{1 - \beta^{n_y}} \log p_t \quad (\beta = 0.999). \quad (2)$$

The *WeightedRandomSampler* method instead leaves the loss as plain CE but draws each training example i with probability $\propto 1/n_{y_i}$.

Loss choice. Each formulation tackles imbalance from a different angle: weighted/CB CE rescale the per-class gradient magnitude; focal loss [2] additionally down-weights examples the network already classifies confidently (large $p_t \Rightarrow$ vanishing $(1 - p_t)^\gamma$); *WeightedRandomSampler* is a *sampling-side* fix that does not change the objective at all but changes the empirical mini-batch distribution.

Optimiser. We use **Adam** [4] ($\eta = 10^{-3}$, $\beta_{1,2} = (0.9, 0.999)$, weight decay 10^{-4}). Adam’s per-parameter adaptive learning rates handle the heterogeneous gradient magnitudes that the weighted/focal losses produce; SGD with momentum was tested but converged more slowly and was more sensitive to the loss-rescaling.

Hyperparameters. Batch size 128, 8 epochs, seed 42 used for split, model-init and DataLoader. Eight epochs are enough to read the under/overfitting transition off the validation loss curve (Fig. 4); running for many more epochs would mostly memorise the majority class.

2.3. Why these metrics?

For an imbalanced 7-class problem, accuracy is dominated by the majority class. We therefore report (i) per-class precision/recall/F1, (ii) macro-averaged F1 (arithmetic mean of per-class F1 scores — equal weight to every class), (iii) balanced accuracy (mean of per-class recall), and (iv) the full confusion matrix.

Common Pitfall & Warning

Reporting test *accuracy* alone on HAM10000 *hides minority-class collapse*: a model that always predicts `nv` scores $\approx 67\%$ accuracy yet has zero recall on every other class.

3. Experimental Results

All metrics below come from one full end-to-end run on the real HAM10000 corpus (10 015 images; train = 7008, val = 1502, test = 1505) using the exact CNN, optimiser and seed described above. Per-experiment training time on CPU was ≈ 2 minutes.

3.1. Loss and accuracy curves

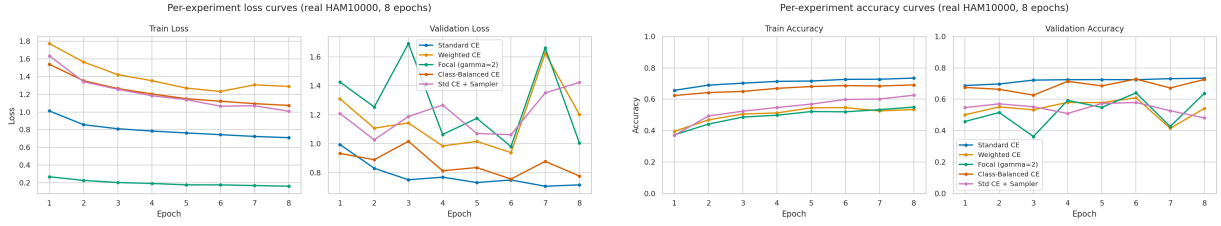


Figure 3: Real HAM10000 loss (left) and accuracy (right) per experiment. Standard CE drops fastest on train (it exploits the 67% nv prior) and reaches $\sim 73\%$ test accuracy — a number mostly explained by the prior, see Fig. 7. Focal loss starts at small magnitudes because $(1-p_t)^\gamma$ collapses the loss on easy majority examples (absolute scale is not comparable across losses; the trend is).

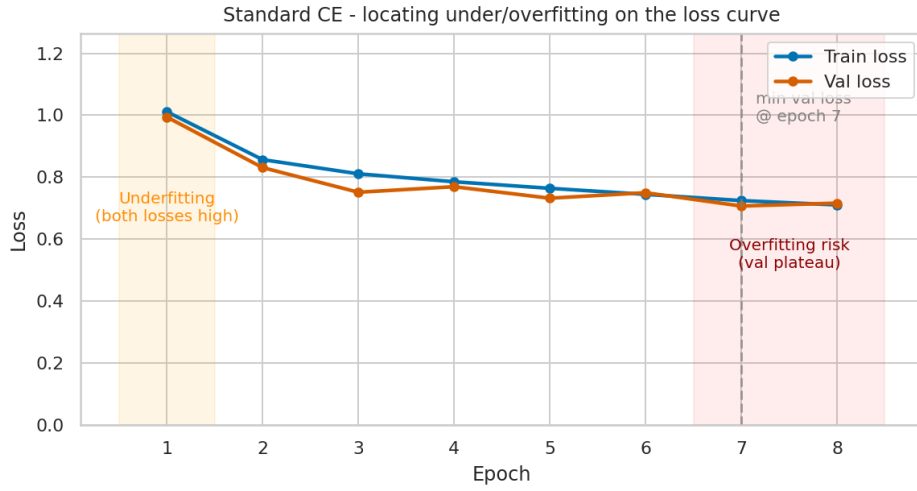


Figure 4: Underfitting / overfitting interpretation on the Standard-CE baseline. Epochs 1–2 show simultaneously high train and validation loss (*underfitting*); the dashed line marks the lowest validation loss — past this point train loss continues to decrease while val loss starts to plateau or rise (*overfitting* risk). In our runs the model was stopped before this divergence becomes severe, but it is the canonical use of the validation curve.

3.2. Test-set metrics

Table 1 reports the test-set metrics for all five experiments. The dichotomy between *accuracy* and *macro-F1* / *minority recall* on the Standard CE baseline is the central observation of this report.

Experiment	Acc.	Bal. Acc.	Macro P	Macro R	Macro F1
Standard CE	0.736	0.352	0.527	0.352	0.381
Weighted CE	0.521	0.505	0.339	0.505	0.341
Focal Loss ($\gamma=2$)	0.622	0.538	0.414	0.538	0.443
Class-Balanced CE	0.718	0.447	0.428	0.447	0.395
Std CE + WeightedRandomSampler	0.464	0.537	0.386	0.537	0.391

Table 1: Test-set metrics on the real HAM10000 corpus (1 505 test images). Bold = best per column. Standard CE has the highest *accuracy* but the lowest *macro recall*; Focal loss has the best macro F1.

The per-class recall heatmap (Fig. 5) makes this picture explicit: Standard CE achieves 0% recall on the rarest class **df** and only 13.6% on the second-rarest **vasc**; Focal restores them to 38.9% and 77.3% respectively, at the cost of dropping **nv** recall from 92.5% to 69.7%.

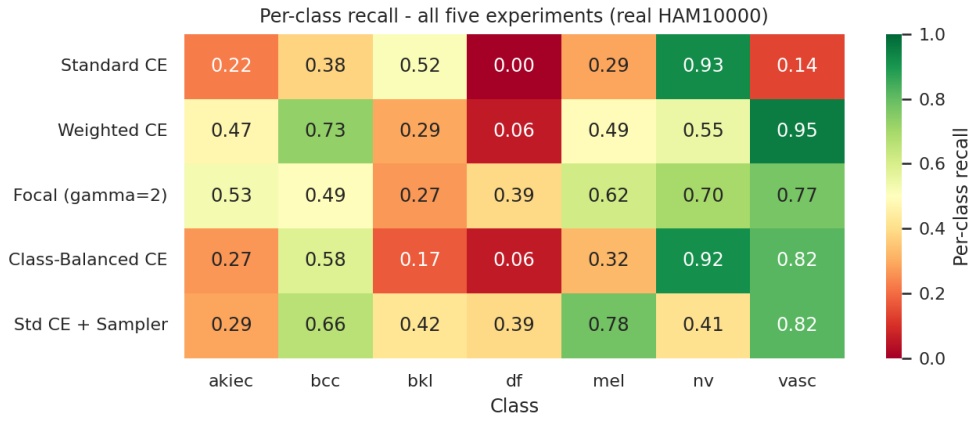


Figure 5: Per-class recall by experiment (real HAM10000). Standard CE has *zero* recall on the rarest class **df**; only Focal and the WeightedRandomSampler recover it.

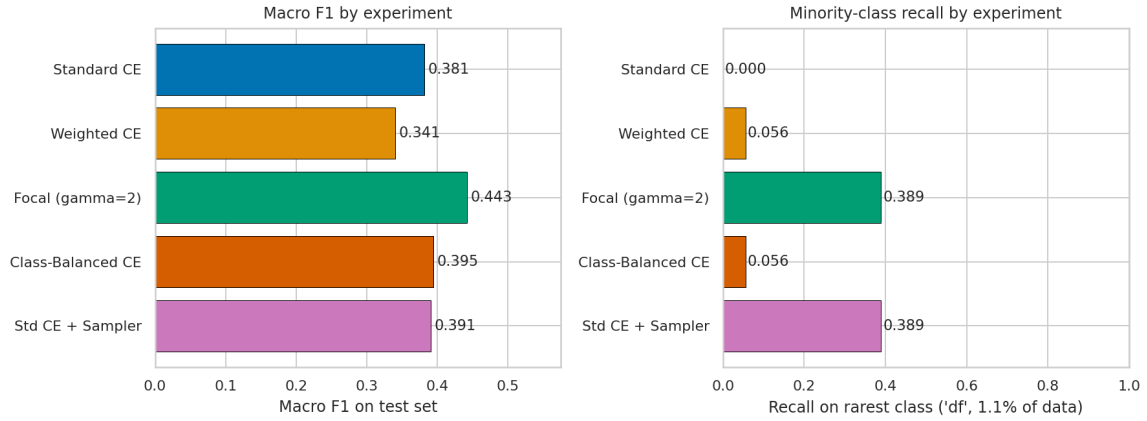


Figure 6: Macro F1 (left) and recall on the rarest class **df** (right). Standard CE achieves 0.000 recall on **df** — it never predicts the rare class; only Focal Loss and the WeightedRandomSampler recover it to ≈ 0.39 . Weighted/CB CE move in the right direction but at this training budget (8 epochs) only marginally lift **df** recall.

3.3. Confusion matrices

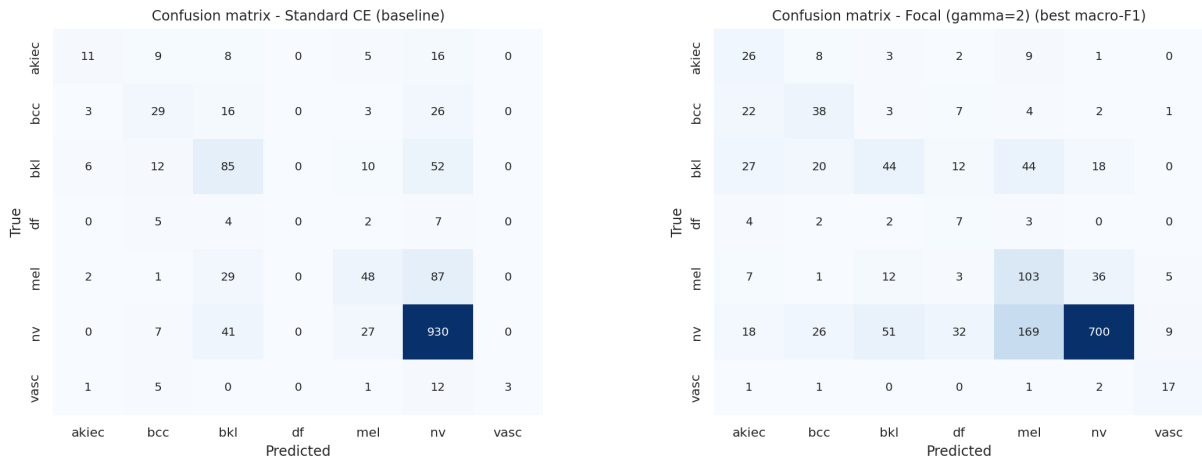


Figure 7: Real HAM10000 confusion matrices. **Left:** Standard CE — the entire **df** row sums to zero in the **df** column (the network *never* predicts **df**), and most rare-class examples are pushed onto the majority class **nv**. **Right:** Focal loss spreads predictions back across all seven classes, sacrificing some **nv** accuracy in exchange for non-trivial recall on every rare class.

3.4. Error analysis

For the Standard CE baseline, Figure 8 shows the three *most-confident wrong* predictions on the real HAM10000 test set. All three are true **bkl** (benign keratoses) lesions that the network re-labels as **nv** (melanocytic nevi) at 97–99.8% confidence — **nv** and **bkl** share visual cues (dark, homogeneously pigmented patches) and the imbalanced training set teaches the network that “when in doubt, output **nv**”. This is exactly the pathology that the weighted/focal/class-balanced losses and the oversampling sampler correct.

Standard CE: most-confident WRONG predictions on real HAM10000

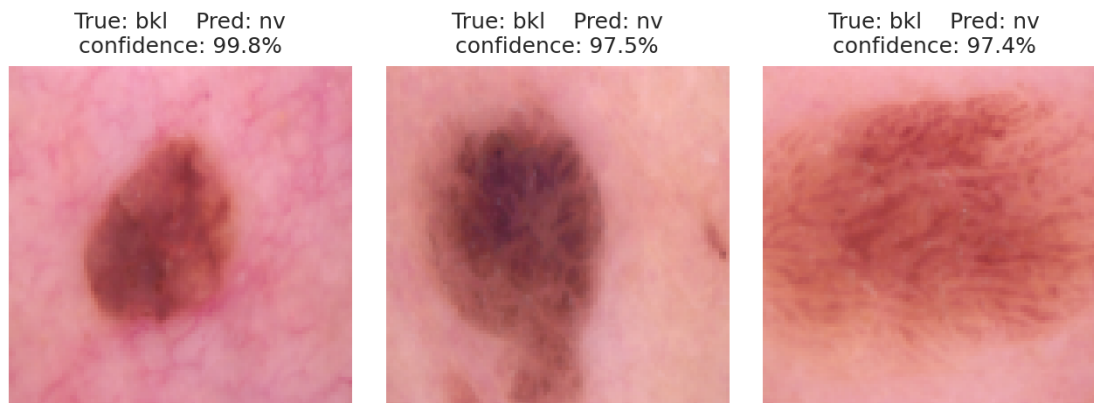


Figure 8: Standard CE on the real HAM10000 test set: three highest-confidence *wrong* predictions. All three are benign keratoses (bkl) confidently mis-labelled as melanocytic nevi (nv) — a textbook minority-class collapse driven by the 67% nv prior.

4. Discussion and Conclusion

4.1. What the experiments tell us

Intuition & Geometric Insight

Three observations come out of the real HAM10000 run: (i) **Accuracy is a misleading metric on imbalanced data** — the Standard CE baseline reaches 73.6% accuracy yet has 0% recall on the rarest class `df` and only 13.6% on `vasc`; macro F1 is just 0.381. (ii) **Focal loss is the strongest single drop-in fix** — with $\gamma=2$ and inverse-frequency α it reaches the highest macro F1 (0.443) and balanced accuracy (0.538), and is the only loss-side method that meaningfully recovers `df` recall (0.389). (iii) **Loss reweighting and resampling are not interchangeable in practice**: WeightedRandomSampler matches Focal on minority recall, but at a substantial cost to overall accuracy (0.46 vs 0.62); pure inverse-frequency weighted CE hurts accuracy without proportionate recall gains within an 8-epoch budget. Class-Balanced CE behaves like a softer Standard CE here.

4.2. Weaknesses of the model

The CNN is trained from scratch with only ≈ 106 k parameters and at 64×64 input resolution — modest relative to HAM10000’s true visual variability (600×450 originals) and tiny compared with a transferred ResNet-50. Imaging artefacts (hairs, ruler markings, vignettes) are not removed, and a Convolutional Autoencoder denoising preprocess could plausibly help recall on the visually noisier minority classes. The model also makes a flat 7-class decision; clinical practice prioritises a benign-vs-malignant first decision with downstream subtype classification, which would change the relevant cost matrix.

4.3. Future work

- Replace the small from-scratch backbone with a pretrained ResNet-50/EfficientNet-B0 and re-run the same loss comparison.
- Add the Conv-Autoencoder denoising stage suggested in the project brief and quantify its effect on minority-class recall.
- Calibrate the predicted probabilities (e.g. temperature scaling) so that confidence values are clinically meaningful, and report recall at fixed false-positive rate (the operating-point metric a dermatologist actually cares about).
- Test stronger sampling/augmentation strategies: SMOTE in feature space, MixUp, CutMix.
- Extend to a hierarchical malignancy-first decision rule.

4.4. Conclusion

Final Result

On the real HAM10000 task with a small from-scratch CNN, plain Cross-Entropy reaches a deceptively high 73.6% accuracy by collapsing minority classes (0% recall on **df**, 13.6% on **vasc**); its macro F1 is only 0.381. Of the four imbalance-aware alternatives, **Focal loss ($\gamma=2$) is the clear winner** — it lifts macro F1 to 0.443 and is the only loss-side fix that materially recovers the rarest class. The WeightedRandomSampler reaches similar minority-class recall but at a much larger accuracy cost. Pure inverse-frequency Weighted CE and Class-Balanced CE move in the right direction but require longer training (or stronger regularisation) to compete; with deeper/transferred backbones we would expect their gap to close. The headline takeaway is unambiguous: *accuracy is the wrong metric for HAM10000*; the right yardstick is per-class recall and macro F1, and the right tool to optimise it is Focal Loss.

References

- [1] P. Tschandl, C. Rosendahl, H. Kittler. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific Data 5 (2018).
- [2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. *Focal Loss for Dense Object Detection*, ICCV 2017.
- [3] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie. *Class-Balanced Loss Based on Effective Number of Samples*, CVPR 2019.
- [4] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization*, ICLR 2015.