

AI-Assisted Diagnosis on HAM10000

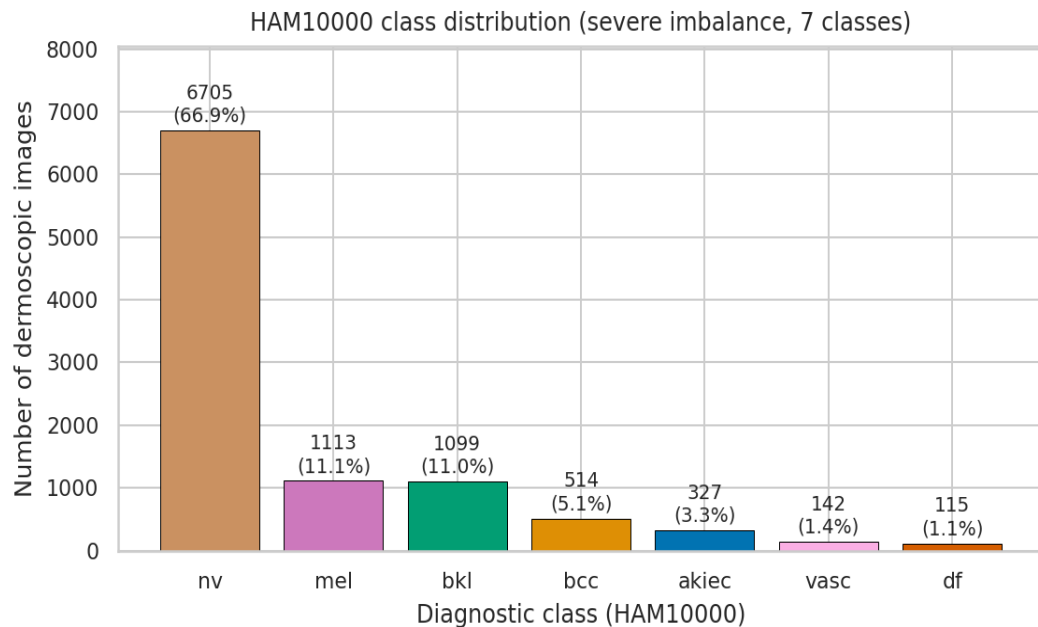
The effect of loss functions and sampling on imbalanced skin-cancer classification

Ömer Kahyaoglu

Student no: 090240741 · kahyaoglu20@itu.edu.tr

What happens when 67% of your data is one class?

HAM10000: 10,015 dermoscopic images, 7 diagnostic classes



67%

Majority class
(melanocytic nevi)

1.1%

Rarest class
(dermatofibroma)

58x

Imbalance ratio
(majority : minority)

What does the pipeline look like?

Kaggle download -> stratified 70/15/15 split -> CNN -> imbalance-aware metrics



CNN architecture (106,391 trainable parameters)

Conv 3->16 + BN + ReLU + MaxPool 2 (16 x 32 x 32)

Conv 16->32 + BN + ReLU + MaxPool 2 (32 x 16 x 16)

Conv 32->64 + BN + ReLU + MaxPool 2 (64 x 8 x 8)

Conv 64->128 + BN + ReLU + MaxPool 2 (128 x 4 x 4)

GlobalAvgPool -> **Dropout(0.4)** -> **FC 128->64** -> **ReLU** -> **Dropout(0.4)** -> **FC 64->7**

What if we change the loss function or sampling?

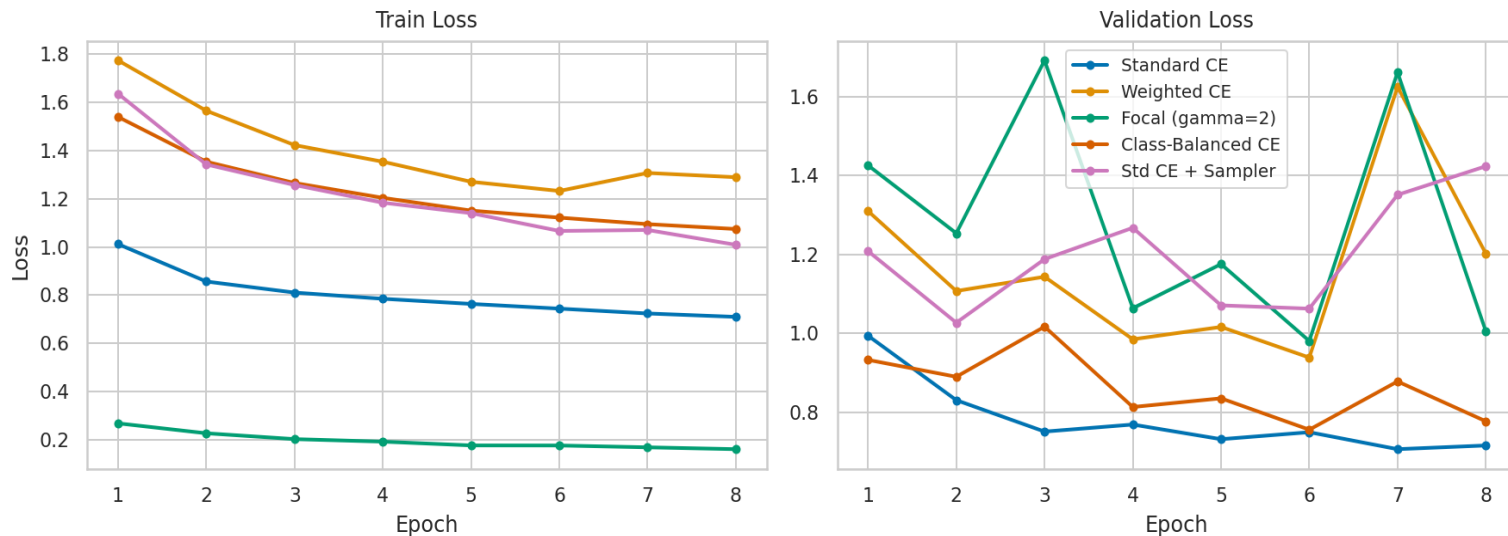
Five configurations - same model, same optimiser, same seed

Experiment	Idea
Standard CE	Baseline; no rebalancing
Weighted CE	Upweight rare classes
Focal Loss	De-emphasise easy examples
Class-Balanced CE	Effective number of samples
Std CE + Sampler	Sampling-side fix

What if we plot the loss across epochs?

Train (left) and validation (right) loss for all five configurations

Per-experiment loss curves (real HAM10000, 8 epochs)



Reading the curves

Standard CE:

lowest train + val loss - by exploiting the 67% nv prior

Focal:

smallest absolute scale because $(1-p_t)^\gamma$ collapses easy examples

Rewighted:

noisier val loss; class weights amplify minority gradient noise

What if we look beyond accuracy?

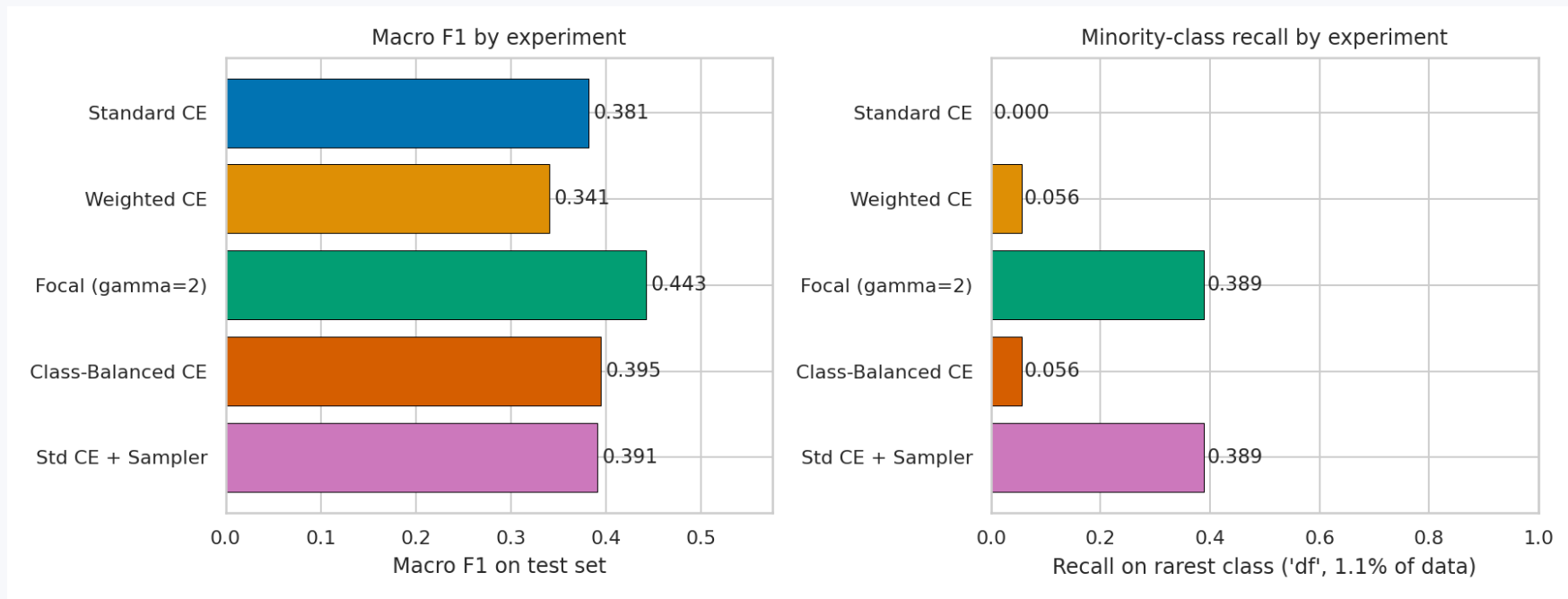
Standard CE: 73.6% accuracy AND 0% recall on the rarest class

Experiment	Accuracy	Bal. Acc.	Macro F1	'df' Recall
Standard CE	0.736	0.352	0.381	0.000
Weighted CE	0.521	0.505	0.341	0.056
Focal Loss (gamma=2)	0.622	0.538	0.443	0.389
Class-Balanced CE	0.718	0.447	0.395	0.056
Std CE + Sampler	0.464	0.537	0.391	0.389

Macro F1 = arithmetic mean of per-class F1 - equal weight to every class. Balanced Acc = mean of per-class recall.

What if we visualise rare-class recall?

Standard CE never predicts df; only Focal and the Sampler meaningfully recover it



What if we look at the confusion matrix?

All three test-set 'df' samples collapse onto 'nv' in the baseline

Confusion matrix - Standard CE (baseline)							
True	akiec	bcc	bkl	df	mel	nv	vasc
	11	9	8	0	5	16	0
	3	29	16	0	3	26	0
	6	12	85	0	10	52	0
	0	5	4	0	2	7	0
	2	1	29	0	48	87	0
	0	7	41	0	27	930	0
	1	5	0	0	1	12	3
		Predicted					

Standard CE - df column entirely empty

Confusion matrix - Focal (gamma=2) (best macro-F1)							
True	akiec	bcc	bkl	df	mel	nv	vasc
	26	8	3	2	9	1	0
	22	38	3	7	4	2	1
	27	20	44	12	44	18	0
	4	2	2	7	3	0	0
	7	1	12	3	103	36	5
	18	26	51	32	169	700	9
	1	1	0	0	1	2	17
		Predicted					

Focal Loss (best macro F1 = 0.443)

What if we look at where the baseline fails?

Three most-confident WRONG predictions of Standard CE - all true 'bkl' lesions called 'nv' at 97-99.8% confidence

Standard CE: most-confident WRONG predictions on real HAM10000

True: bkl Pred: nv
confidence: 99.8%



True: bkl Pred: nv
confidence: 97.5%



True: bkl Pred: nv
confidence: 97.4%



Network has effectively learnt: 'when in doubt, predict the majority class'. Loss reweighting and oversampling fix exactly this.

What did we learn?

Three findings transferable to any imbalanced classification task

Accuracy is misleading

Standard CE reaches 73.6% accuracy yet 0% recall on the rarest class df, 13.6% on vasc. Always report macro F1 + per-class recall on imbalanced data.

Focal loss wins decisively

Focal reaches macro F1 = 0.443 vs 0.381 for Standard CE; the only loss-side fix that recovers df recall (0% -> 38.9%).

Loss vs sampling: not a tie

WeightedRandomSampler matches Focal on minority recall (38.9%) but at a much larger accuracy cost (46% vs 62%). Pure Weighted CE under-performs both within an 8-epoch budget.

Thank you - Questions?

What if we kept iterating? Future work below.

- Replace the from-scratch backbone with pretrained ResNet-50 / EfficientNet-B0
- Add the Convolutional Autoencoder denoising stage (hairs, ruler markings, vignettes)
- Calibrate predicted probabilities (temperature scaling) for clinically meaningful confidence
- Test stronger sampling/augmentation: SMOTE-in-feature-space, MixUp, CutMix
- Hierarchical decision: malignant-vs-benign first, subtype second