

Interpretable Skin Lesion Classification via Concept Bottleneck Models on the MILK10k Dataset

Pradeep Banavara
Independent Researcher

Abstract

We present a Concept Bottleneck Model (CBM) for skin lesion diagnosis on the MILK10k benchmark, which comprises 5,240 lesions with paired clinical close-up and dermoscopic images across 11 diagnostic categories. Our hybrid CBM architecture uses a shared DINOv2 ViT-L/14 backbone to extract features from both image modalities, predicts 7 MONET dermoscopic concepts as an interpretable bottleneck, and combines concept predictions with a learned residual vector for final diagnosis classification. Through systematic iterative development -- progressing from a strict CBM (F1: 0.23) through architectural changes, loss rebalancing, data augmentation, and class weighting -- we achieve a best macro F1 of 0.5129 on the validation set. We describe our design decisions, failure modes encountered, and lessons learned in training CBMs on highly imbalanced, small-scale medical imaging datasets. All experiments were conducted on an NVIDIA GH200 using bf16 mixed precision.

1. Introduction

Concept Bottleneck Models (CBMs) [1] offer an interpretable alternative to black-box classifiers by forcing predictions through a set of human-understandable intermediate concepts. In dermatology, where clinical decision-making already relies on structured dermoscopic features (e.g., pigment network patterns, vascular structures, regression areas), CBMs are a natural fit: the model first predicts clinically meaningful attributes, then uses these to classify the lesion. A key design choice in our architecture is that the concept head operates on a fused representation that combines both clinical and dermoscopic views of the same lesion, rather than on raw backbone features from a single modality. This mirrors dermatological practice, where clinicians integrate information from both views before assessing dermoscopic attributes. The MILK10k dataset presents a challenging benchmark: 5,240 lesions with paired clinical close-up and dermoscopic images (10,480 total), annotated with 7 MONET dermoscopic concept scores and 11 diagnostic labels. The class distribution is heavily imbalanced, ranging from 2,522 samples for basal cell carcinoma (BCC) to just 9 for other malignancies (MAL_OTH). This paper describes our CBM approach, the iterative development process, and the design decisions that led to our final model.

2. Method

2.1 Architecture

Our model employs a shared DINOv2 ViT-L/14 backbone (304M parameters) [2] to process both image modalities. The CLS token embeddings from clinical and dermoscopic images (each 1024-d) are concatenated and projected through a fusion MLP to a 512-dimensional representation. This fused representation feeds three parallel heads:

- Concept head: A two-layer MLP (512->256->7) with sigmoid activation, predicting 7 MONET dermoscopic concept scores in [0, 1].
- Residual head (hybrid only): A linear projection (512->16) capturing non-concept information that bypasses the interpretable bottleneck.
- Classification head: A two-layer MLP (23->64->11) taking the concatenation of 7 concept predictions and 16 residual dimensions as input.

We explored three CBM variants: strict (classification sees only concepts), hybrid (concepts + residual), and baseline (no bottleneck). The hybrid variant was selected as the final architecture based on validation performance.

2.2 Two-Phase Training

Training proceeds in two phases to balance feature adaptation with head optimization:

Phase 1 (10 epochs): The DINOv2 backbone is frozen. Only the fusion MLP, concept head, residual head, and classification head are trained with AdamW (lr=1e-3, weight decay=0.05). This allows the lightweight heads to calibrate against the pretrained backbone's feature space.

Phase 2 (up to 50 epochs): The full model is fine-tuned with differential learning rates: backbone at 2e-6, heads at 1e-4. A cosine annealing schedule with 5% linear warmup (stepped per batch) controls the learning rate. Early stopping with patience 7 (based on macro F1) prevents overfitting.

2.3 Loss Function

The joint loss combines concept supervision and classification: $L = \alpha * \text{MSE}(\text{concepts}) + \beta * \text{CE}(\text{classification})$, where $\alpha=5.0$ and $\beta=1.0$. The elevated concept weight was necessary to prevent classification gradients from dominating during Phase 2 fine-tuning (Section 3.2). The cross-entropy loss uses inverse-frequency class weighting to address the severe class imbalance: each class weight is proportional to $1/n_k$, normalized to sum to the number of classes. This amplifies the gradient contribution of rare classes (e.g., MAL_OTH weight: 5.95 vs. BCC weight: 0.02).

2.4 Data Pipeline

Patient-level stratified splitting (80/20) ensures no lesion appears in both train and validation sets. The training augmentation pipeline applies: RandomResizedCrop (224, scale 0.7-1.0), random horizontal and vertical flips, RandomAffine (rotation +/-30 deg, translation 10%, scale

0.9-1.1, shear ± 10 deg), ColorJitter (brightness 0.3, contrast 0.3, saturation 0.2, hue 0.05), RandAugment (2 ops, magnitude 9), and RandomErasing ($p=0.25$). Validation uses Resize(224) followed by CenterCrop(224). All images are normalized to ImageNet statistics.

2.5 Infrastructure

All experiments ran on an NVIDIA GH200 480GB (94.5 GB HBM3 GPU memory) with bfloat16 mixed-precision training. The GH200's native bf16 support (same exponent range as fp32) eliminates the need for gradient scaling, simplifying the AMP pipeline. Batch size was 64 with 8 data-loading workers.

3. Iterative Development and Ablations

Our development followed a systematic trajectory, with each iteration targeting a specific bottleneck. Table 1 summarizes the key configurations and their results.

Table 1: Iterative development trajectory. Macro F1 on validation set.

#	Key Change	Val F1	Outcome
1	Strict CBM, scheduler bug (epoch-level)	0.23	LR never exceeded 30% of target
2	Strict CBM, scheduler fixed (batch-level)	0.27	Moderate improvement
3	Switch to hybrid variant	0.29	Residual path helps classification
4	concept_weight 1->5, backbone LR $1e-5 \rightarrow 2e-6$	0.47	Major gain: concepts preserved
5	Strong regularization (drop=0.5, wd=0.1, ls=0.1)	0.24	FAILURE: severe underfitting
6	Light regularization (drop=0.2, no ls)	0.47	Recovered baseline
7	Data augmentation (RandAug, Affine, Erasing)	0.48	Better generalization
8	Inverse-frequency class weighting	0.50	Rare classes improved
9	WeightedRandomSampler oversampling	0.31	FAILURE: BCC collapsed to F1=0
10	Phase 2 extended to 50 epochs	0.51	Best result; early stop epoch 37

3.1 Strict vs. Hybrid CBM

The strict CBM achieved only 0.23-0.27 macro F1, constrained by the 7-dimensional bottleneck. With only 7 concept features visible to the classification head, discriminating among 11 classes (some with <100 samples) proved insufficient. The hybrid variant adds a 16-dimensional learned residual, providing the classifier with 23 total input features. This nearly doubled performance, confirming that not all diagnostic information is captured by the MONET concept annotations.

3.2 Concept Loss Rebalancing

A critical failure mode emerged during Phase 2 fine-tuning: concept MSE loss (~ 0.08) was an order of magnitude smaller than classification CE loss (~ 1.0). When the backbone was unfrozen, classification gradients dominated, causing the concept head to regress. Increasing the concept loss weight from 1.0 to 5.0 and reducing the backbone learning rate from $1e-5$ to $2e-6$ resolved this, yielding the single largest F1 improvement ($+0.20$). This highlights a known challenge in multi-task learning where loss magnitude imbalances can destabilize shared representations.

3.3 Regularization: Less is More

Counter-intuitively, aggressive regularization severely hurt performance on this small dataset. Increasing dropout from 0.2 to 0.5 with weight decay 0.1 and label smoothing 0.1 caused the model to underfit, dropping F1 from 0.47 to 0.24. The effective strategy was minimal regularization (dropout 0.2, weight decay 0.05, no label smoothing) combined with data augmentation for implicit regularization. The key insight: with only 4,192 training lesions, the model needs maximum capacity to learn discriminative features before any regularization is applied.

3.4 Class Imbalance Handling

Two approaches were evaluated. Inverse-frequency class weighting in the CE loss improved rare-class F1 substantially (DF: 0.14->0.50, INF: 0->0.22, VASC: 0.55->0.67) without degrading dominant classes. WeightedRandomSampler for batch-level oversampling was too aggressive: equalizing class frequencies in each batch caused BCC (the largest class) to collapse from F1 0.89 to 0.00, with severe overfitting (train loss 0.05 vs. val loss 2.4). Loss-level weighting proved far more stable than sampling-level balancing for this degree of imbalance.

4. Results

The final model (hybrid CBM, 50 Phase 2 epochs, class-weighted CE, full augmentation) achieved 0.5129 macro F1 on the validation set, early-stopping at epoch 37. Table 2 reports per-class F1 at the best checkpoint.

Table 2: Per-class validation F1 at best checkpoint (macro F1 = 0.5129).

Class	F1	Train N	Class	F1	Train N
BCC	0.859	2,522	NV	0.785	745
SCCKA	0.685	473	VASC	0.636	47
MEL	0.615	449	AKIEC	0.544	302

BKL	0.477	543	BEN_OTH	0.353	44
DF	0.400	52	INF	0.105	50
MAL_OTH	0.000	9			

Performance correlates with class frequency, with a notable exception: VASC (47 samples) achieves 0.636 F1, likely due to distinctive vascular morphology. MAL_OTH (9 samples) remains unresolvable at this dataset scale. The concept head achieves MSE of ~ 0.018 and positive Pearson correlations across all 7 MONET concepts, confirming meaningful intermediate representations.

5. Discussion and Limitations

Interpretability trade-off. The hybrid CBM sacrifices full interpretability for performance. The 16-dimensional residual path is opaque, meaning only the 7-concept pathway is directly interpretable. A strict CBM maintains full transparency but at a significant performance cost (-0.24 macro F1). For clinical deployment, the hybrid variant offers a practical compromise: clinicians can inspect predicted concept values for a "reason check" while benefiting from the residual's added discriminative power.

Dataset limitations. With 5,240 lesions, MILK10k is small by modern deep learning standards. The extreme imbalance (281:1 ratio between BCC and MAL_OTH) makes certain classes fundamentally unlearnable. The 7 MONET concepts, while clinically meaningful, may not capture all diagnostic features needed for 11-class discrimination -- evidenced by the strict CBM's poor performance.

Submission format. Our final submission uses sigmoid activation (independent per-class probabilities) rather than softmax, allowing the model to express uncertainty across multiple diagnostic categories simultaneously. This is clinically appropriate, as lesions may exhibit features of multiple conditions.

6. Conclusion

We demonstrated that a hybrid CBM with a DINOv2 backbone can achieve competitive classification performance while maintaining partial interpretability through dermoscopic concept predictions. The key lessons from our iterative development are: (1) loss magnitude balancing is critical in multi-task CBMs, (2) minimal regularization with aggressive augmentation outperforms strong regularization on small medical datasets, and (3) loss-level class weighting is more stable than sampling-level balancing under extreme imbalance. Our final model achieves 0.5129 macro F1 with meaningful concept predictions across all 7 MONET features.

References

- [1] Koh, P.W., Nguyen, T., Tang, Y.S., et al. Concept bottleneck models. ICML, 2020, pp. 5338-5348.
- [2] Oquab, M., Darcet, T., Moutakanni, T., et al. DINOv2: Learning robust visual features without supervision. TMLR, 2024.
- [3] Bissoto, A., Valle, E., and Avila, S. Debiasing skin lesion datasets and models? Not so fast. CVPR Workshops, 2020.
- [4] Tschandl, P., Rosendahl, C., and Kittler, H. The HAM10000 dataset. Scientific Data, 5, 180161, 2018.