Supplementary Materials for "Integrating Clinical Knowledge into Concept Bottleneck Models"

Table 4. Classifier layer names and their architectures in PyTorch syntax.

Classifier Name	Architecture
Linear	`nn.Linear(n_input, n_classes)`
MLP(20)	`nn.Linear(n_input, 20)` → `nn.ReLU()` → `nn.Linear(20, n_classes)`
MLP(128)	`nn.Linear(n_input, 128)` → `nn.ReLU()` → `nn.Linear(128, n_classes)`

Table 5. Implementation details. We run the same code three times with different seeds and report 95% confidence intervals.

Dataset	Backbone +Classifier	Optimizer	Batch size	Epoch	LR	Weight decay	φ	LS	λ
WBC	Vgg16+Linear	AdamW	64	30	0.0001	0.01	1	0.3	1
	Vgg16+MLP(20)							0.05	1
	Vgg16+MLP(128)							0.3	1
	ViT-B/16+Linear	AdamW	64	30	0.0001	0.01	1	0.3	3
	ViT-B/16+MLP(20)							0.1	1
	ViT-B/16+MLP(128)							0.3	1
Skin	Vgg16+Linear	AdamW	64	30	0.0001	0.01	1	0.1	5
	Vgg16+MLP(20)							0.1	1
	Vgg16+MLP(128)							0	5
	ViT-B/16+Linear	AdamW	64	30	0.0001	0.01	1	0.5	5
	ViT-B/16+MLP(20)							0	2
	ViT-B/16+MLP(128)							0.1	1

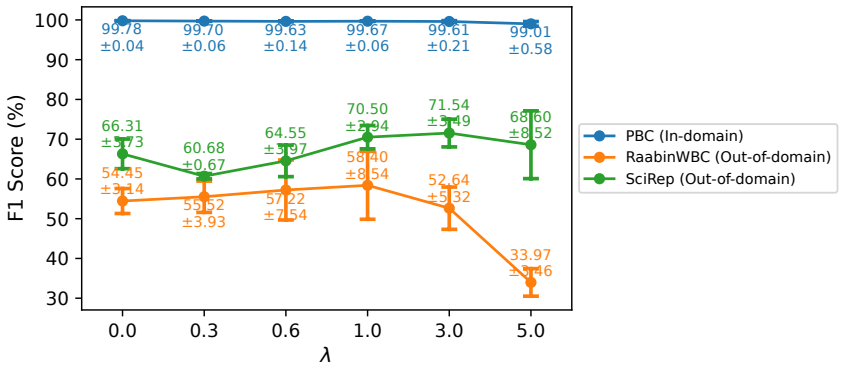

Fig. 4. The effect of λ in loss function for WBC classification, using VGG + MLP(128).

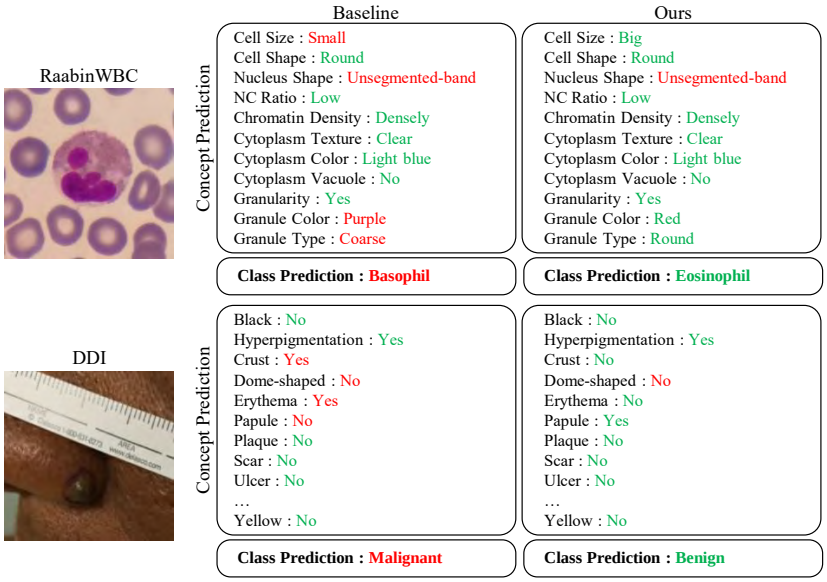

Fig. 5. Qualitative results on the out-of-domain WBC and skin datasets demonstrate that our method improves concept predictions, leading to correct class predictions.

Final version

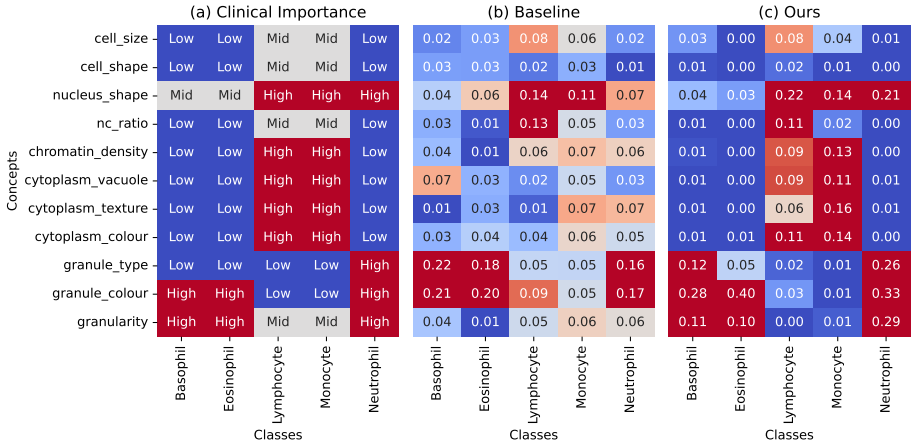

Fig. 6. (a) The importance level of the concepts focused on by a pathologist in classifying WBC images into their types. (b), (c) Concept importance learned by the models (Baseline vs. Ours) for WBC classification. The concept importance learned by (c) our model is better aligned with the clinical importance.