SUPPLEMENTARY MATERIAL: TOWARDS CONCEPT-BASED INTERPRETABILITY OF MELANOMA DIAGNOSIS USING VISION-LANGUAGE MODELS

Cristiano Patrício[1,3], *Luis F. Teixeira*[2,3], *João C. Neves*[1]

[1]Universidade da Beira Interior and NOVA LINCS,
[2]Faculdade de Engenharia da Universidade do Porto, [3]INESC TEC

1. MELANOMA DIAGNOSIS WITH CONCEPT-BASED EXPLANATIONS

Figure 1 displays images classified as "Melanoma" and "Nevi" using our method. The weights of the linear classifier were recovered from a linear classifier specifically trained on dermoscopy images for melanoma diagnosis [1]. These weights correspond to the importance of each concept to the target label. Most of them are related with the ABCDEs of melanoma. An examination of the concept coefficients reveals that positive weights are assigned to concepts exhibiting strong correlations with melanoma.

2. DERMOSCOPIC CONCEPTS

Table 6 provides the results generated by ChatGPT based on the designated prompt. The aim was to create a set of m textual descriptions for specific dermoscopic concepts, as indicated in the "Concept" column of Table 6. The chosen prompt, "According to published literature in dermatology, which phrases best describe a skin image containing `concept`?", was employed to obtain a total of five descriptions for each concept c. Subsequently, we encoded each of these descriptions using the text encoder of the CLIP model.

3. LINEAR PROBE MODELS

In our experiments, we determine the best L2 regularization strength λ using a hyperparameter search on the validation splits of each dataset over the range between 10^{-5} and 10^{0}, with 960 spaced steps. All models were trained on an NVIDIA GTX TITAN X GPU. Table 1 reports the detailed results for the evaluated linear probe models in terms of Balanced Accuracy.

4. NETWORK PARAMETERS

The network parameters used during training are specified in Table 2.

Model		Dataset		
		PH^2 [2]	Derm7pt [3]	ISIC [4]
CLIP [5]	RN50	68.9	73.8	62.4
	RN101	69.2	74.5	59.2
	ViT-B/16	84.9	78.9	63.2
	ViT-B/32	83.9	80.5	61.8
	ViT-L/14	81.6	**83.2**	62.7
	RN50x16	78.1	76.8	61.3
MONET [1]	ViT-L/14	67.2	82.4	**66.2**

Table 1: Linear probe performance (Balanced Accuracy %) of various pre-trained models over 3 skin image datasets.

Parameter	RN50	RN101	ViT-B/16	ViT-B/32	ViT-L/14	RN50x16
Batch size	32					
Learning rate	1e-5					
Optimizer	AdamW [6]					
Weight decay	1e-3					
Temperature	1.0					
Dropout	0.2					
Patience	1					
Factor	0.8					
Epochs	100					
Image embedding	1024	512	512	512	768	768
Text embedding	1024	512	512	512	768	768
Projection size	1024	512	512	512	768	768

Table 2: Network parameters.

5. STATISTICAL TESTING

To assess statistical significance, we conduct hypothesis testing using the `scipy.stats` library. We utilize a t-test with a significance level of $\alpha = 0.05$ across all datasets to examine the null hypothesis that two related samples share the same average Balanced Accuracy. Specifically, we compare Balanced Accuracy values for each fold (in the case of PH^2) or each run (Derm7pt and ISIC 2018) between two different strategies within the same model architecture. This leads to three comparisons: i) Baseline vs. CBM; ii) Baseline vs. GPT+CBM; and iii) CBM vs. GPT+CBM.

The Balanced Accuracy results for each fold (PH^2 dataset), and each run (Derm7pt, ISIC 2018) across different strategies are presented in Tables 3, 4, and 5, respectively.

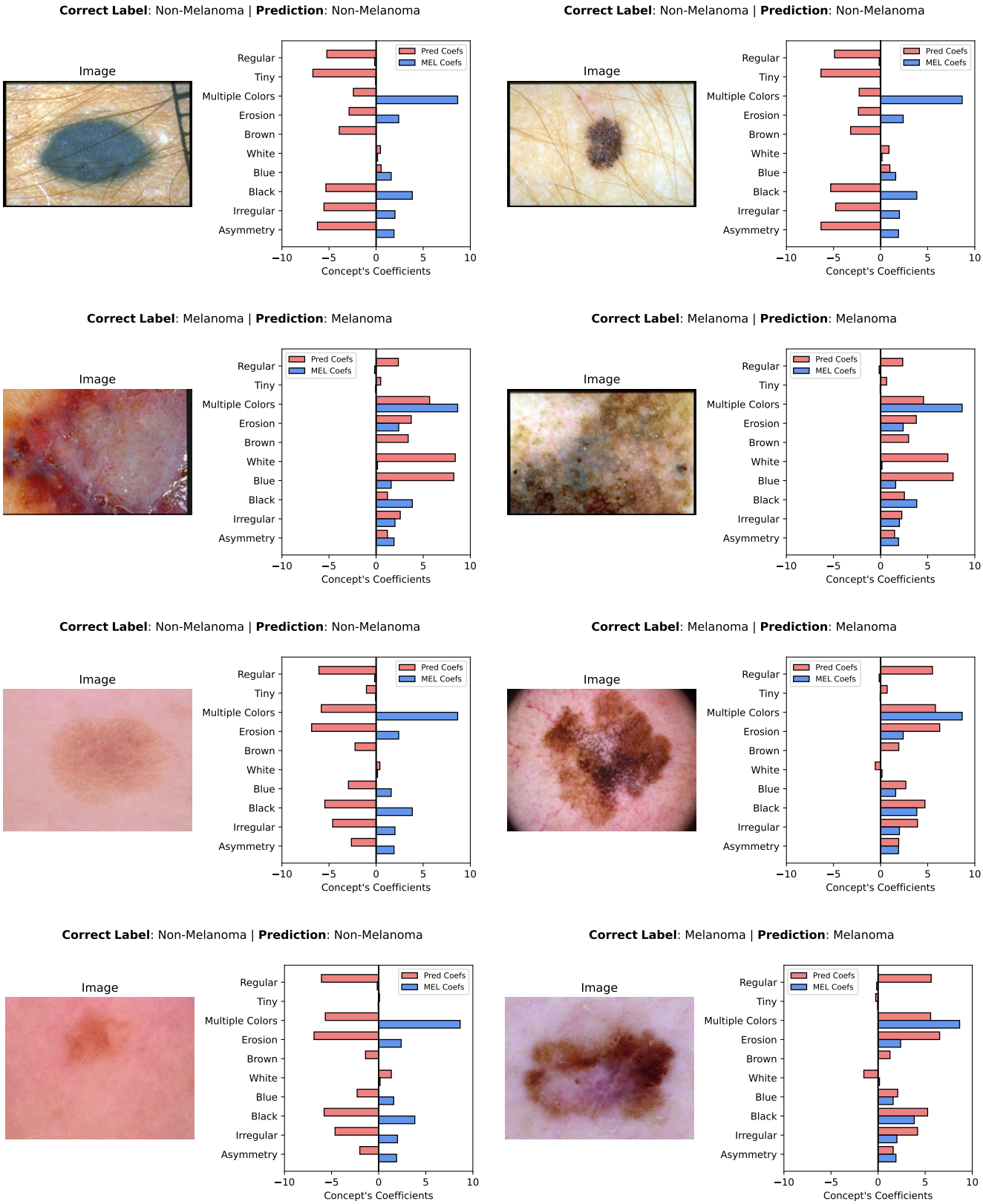

Fig. 1: Examples of ”Melanoma” and ”Nevi” accompanied with the predicted dermoscopic concepts.

	Model		PH^2					
			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	**Mean** ± **Std**
Baseline	CLIP [5]	RN50	58.60	75.98	69.05	69.12	81.17	70.78 ± 7.60
		RN101	85.66	80.39	67.86	69.12	85.28	**77.66** ± 7.73
		ViT-B/16	65.05	59.31	55.36	67.16	75.76	64.53 ± 6.99
		ViT-B/32	51.43	60.78	48.81	45.10	54.98	52.22 ± 5.37
		ViT-L/14	73.66	74.51	69.05	86.76	83.33	<u>77.46</u> ± 6.56
		RN50x16	66.49	79.90	64.29	68.14	71.65	70.09 ± 5.46
	MONET [1]	ViT-L/14	48.03	54.90	43.45	45.10	69.48	52.19 ± 9.49
CBM	CLIP [5]	RN50	63.98	72.55	64.29	76.96	82.25	72.01 ± 7.13
		RN101	67.20	42.65	66.67	56.86	70.13	60.70 ± 10.07
		ViT-B/16	71.33	53.92	61.31	64.71	85.28	67.31 ± 10.59
		ViT-B/32	75.27	51.96	66.07	76.96	70.56	68.16 ± 8.95
		ViT-L/14	82.44	74.51	80.36	78.43	82.25	**79.60** ± 2.93
		RN50x16	81.72	78.92	82.14	74.51	78.14	<u>79.09</u> ± 2.76
	MONET [1]	ViT-L/14	63.44	59.31	64.29	58.82	67.97	62.77 ± 3.38
GPT+CBM	CLIP [5]	RN50	65.77	70.59	81.55	78.92	82.68	<u>75.90</u>± 6.60
		RN101	61.65	47.55	62.50	58.33	70.13	60.03 ± 7.34
		ViT-B/16	80.11	78.92	73.21	74.02	86.36	**78.53** ± 4.75
		ViT-B/32	61.83	57.84	60.71	71.57	64.50	63.29 ± 4.66
		ViT-L/14	69.00	63.73	57.74	69.61	84.85	68.98 ± 9.01
		RN50x16	72.22	78.43	69.64	66.67	69.05	71.20 ± 4.02
	MONET [1]	ViT-L/14	42.65	48.04	47.02	78.92	64.50	56.23 ± 13.56

Table 3: Evaluation results (in BACC %) of the different classification strategies (Baseline, CBM and GPT+CBM) on PH^2 dataset under 5-fold evaluation. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

	Model		Derm7pt				
			Run 1	Run 2	Run 3	Run 4	**Mean** ± **Std**
Baseline	CLIP [5]	RN50	73.09	77.27	70.73	73.17	73.57 ± 2.35
		RN101	75.18	76.82	67.60	71.57	72.79 ± 3.55
		ViT-B/16	75.71	77.62	78.91	78.80	**77.76** ± 1.29
		ViT-B/32	75.79	72.68	74.50	72.98	<u>73.99</u> ± 1.25
		ViT-L/14	74.01	73.13	71.38	72.52	72.76 ± 0.96
		RN50x16	68.57	72.52	73.74	77.27	73.02 ± 3.11
	MONET [1]	ViT-L/14			56.60		
CBM	CLIP [5]	RN50	73.66	75.44	70.20	71.84	72.78 ± 1.96
		RN101	72.52	78.80	65.75	70.80	71.97 ± 4.66
		ViT-B/16	75.38	77.01	76.94	76.66	**76.50** ± 0.66
		ViT-B/32	74.16	75.04	71.46	74.77	<u>73.85</u> ± 1.42
		ViT-L/14	73.62	72.26	72.79	73.36	73.01 ± 0.53
		RN50x16	71.76	73.74	74.00	72.56	73.02 ± 0.90
	MONET [1]	ViT-L/14			62.20		
GPT+CBM	CLIP [5]	RN50	72.90	72.47	70.58	71.53	71.87 ± 0.90
		RN101	72.21	78.38	69.28	71.79	72.92 ± 3.35
		ViT-B/16	74.76	77.81	76.86	76.43	**76.47** ± 1.10
		ViT-B/32	74.73	74.50	71.92	76.33	<u>74.37</u> ± 1.58
		ViT-L/14	74.12	72.52	72.60	73.93	73.29 ± 0.73
		RN50x16	72.33	73.85	74.92	74.00	73.78 ± 0.93
	MONET [1]	ViT-L/14			59.00		

Table 4: Evaluation results (in BACC %) of the different classification strategies (Baseline, CBM and GPT+CBM) on Derm7pt dataset over 4 runs. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

	Model		ISIC 2018				
			Run 1	Run 2	Run 3	Run 4	**Mean** ± **Std**
Baseline	CLIP [5]	RN50	65.55	65.30	65.22	65.11	**65.30** ± 0.16
		RN101	61.06	61.00	62.01	61.50	61.39 ± 0.40
		ViT-B/16	63.80	63.40	64.53	63.40	63.78 ± 0.46
		ViT-B/32	62.55	62.70	61.89	62.41	62.39 ± 0.30
		ViT-L/14	62.56	63.00	62.96	63.69	63.05 ± 0.41
		RN50x16	63.86	63.70	63.57	64.33	<u>63.87</u> ± 0.29
	MONET [1]	ViT-L/14	60.90				
CBM	CLIP [5]	RN50	69.37	71.20	71.27	70.91	70.69 ± 0.77
		RN101	62.92	66.10	62.56	64.79	64.09 ± 1.44
		ViT-B/16	74.76	74.00	72.78	74.39	**73.98** ± 0.75
		ViT-B/32	70.40	71.00	69.24	71.84	70.62 ± 0.95
		ViT-L/14	67.48	67.60	67.90	67.44	67.60 ± 0.18
		RN50x16	72.57	72.40	72.91	72.18	<u>72.51</u> ± 0.27
	MONET [1]	ViT-L/14	64.00				
GPT+CBM	CLIP [5]	RN50	71.28	71.30	73.34	72.96	72.22 ± 0.94
		RN101	66.53	65.60	66.45	67.37	66.49 ± 0.63
		ViT-B/16	73.66	73.00	71.74	73.49	**72.97** ± 0.75
		ViT-B/32	69.74	70.20	69.87	70.26	70.02 ± 0.22
		ViT-L/14	65.91	66.40	66.39	66.58	66.32 ± 0.25
		RN50x16	72.73	72.90	72.59	72.39	<u>72.65</u> ± 0.19
	MONET [1]	ViT-L/14	65.90				

Table 5: Evaluation results (in BACC %) of the different classification strategies (Baseline, CBM and GPT+CBM) on ISIC 2018 dataset over 4 runs. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

6. REFERENCES

[1] Chanwoo Kim, Soham Uday Gadgil, Alex J DeGrave, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee, "Fostering transparent medical image AI via an image-text foundation model grounded in medical literature," *medRxiv*, pp. 2023–06, 2023.

[2] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira, "PH2 - A Dermoscopic Image Database for Research and Benchmarking," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5437–5440.

[3] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning Transferable Visual Models from Natural Language Supervision," in *ICML*, 2021, pp. 8748–8763.

[6] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Concept	ChatGPT Descriptions
Asymmetry	an asymmetric shape with one half not mirroring the other half. asymmetrical distribution of pigmentation. irregular and non-symmetrical borders. significant asymmetry. asymmetry in the form of dissimilar features on opposite sides of the lesion.
Black	dark or black pigmentation black coloration dark brown to black areas black structures or pigmentation black coloration in the form of concentrated dark areas in the lesion
Brown	brown or dark-brown pigmentation brown coloration brown patches or areas of discoloration brown structures or pigmentation This is dermatoscopy od brown coloration in the form of various shades of brown in the lesion
Blue	blue or blue-gray coloration blue coloration bluish patches or areas of discoloration blue structures or pigmentation blue coloration in the form of bluish hues or tones in the lesion
Erosion	surface ulceration or erosion erosion as a crusted area on the skin ulcerated appearance erosion with exposed underlying tissue erosion in the form of disrupted or absent epidermal structures
Irregular	irregular shapes or outlines irregular distribution of pigmentation poorly defined borders irregular and atypical patterns irregular features in the form of non-uniform characteristics
Multiple Colors	a combination of different colors multiple colorations with a varied and complex appearance a mix of different hues diverse colors and pigmentation multiple coloration in the form of different colored areas within the lesion
Regular	a regular and symmetrical pattern regular and evenly spaced structures uniform arrangement of patterns regular pattern regular pattern in the form of symmetrical and well-defined features within the lesion
White	white or hypopigmented coloration white coloration pale or depigmented patches or areas white structures or depigmentation white coloration in the form of reduced pigmentation in the lesion
Tiny	small and minute structures or shapes tiny shapes characterized by their small size minuscule or small-sized patterns tiny structures or shapes tiny shape in the form of small and discrete features within the lesion

Table 6: Dermocopic concepts and the correspondent generated descriptions by ChatGPT.