Melanoma Breslow thickness classification using ensemble-based knowledge distillation with semi-supervised convolutional neural networks - Supplementary material

Juan P. Dominguez-Morales, *Member, IEEE*, Juan-Carlos Hernández-Rodríguez, Lourdes Duran-Lopez, Julián Conejo-Mir, and Jose-Juan Pereyra-Rodriguez

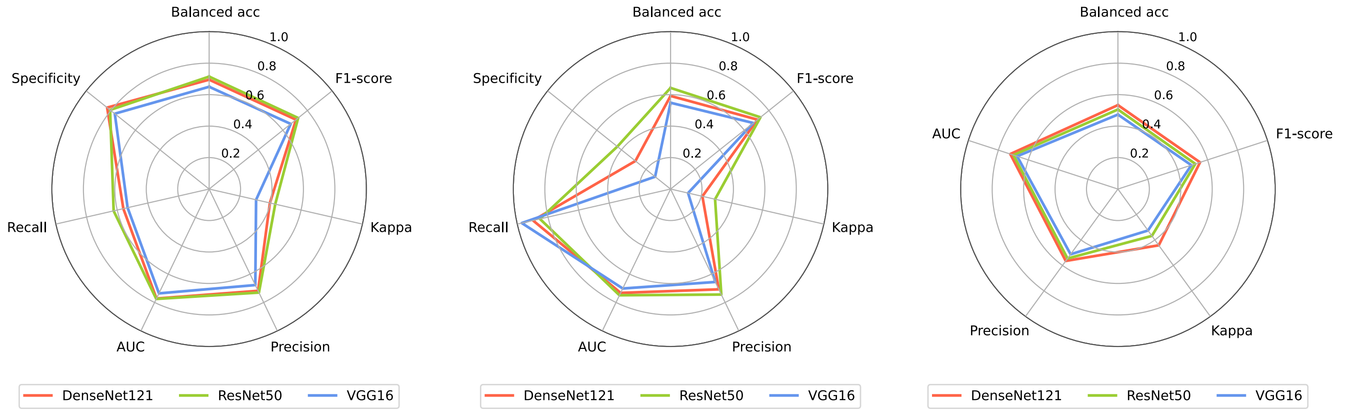
I. Ablation study with different CNN backbones

Different CNN backbones were trained and evaluated in order to compare their performanace on the different tasks considered. These backbones were DenseNet121, ResNet50 and VGG16. Table I summarizes these results. As can be seen, ResNet50 achieves the best performance on Miv vs Mis and on BT < 0.8 vs ≥ 0.8 mm tasks. On the other hand, DenseNet121 reports the highest performance on the Multiclass classification task. Therefore, these were the backbones used in the paper on those specific tasks.

Fig. 1 shows radar plots for each task, model and metric. These represent the same information that can be seen in Table I in a more visual way, which helps understand which model performs better in each of the classification tasks performed.

TABLE I: Summary of the results obtained for each of the training tasks and CNN backbones (VGG16, DenseNet121 and ResNet50), which were trained using supervised learning. Each cell represents the average and standard deviation of the results obtained when evaluating the 5 models in the cross-validation for a specific metric and task.

Task	**Model**	**Balanced acc**	**F1-score**	**Kappa score**	**Precision**	**AUC**	**Recall**	**Specificity**
BT < 0.8 vs ≥ 0.8 mm	DenseNet121	0.6951 ± 0.0685	0.7070 ± 0.0743	0.3966 ± 0.1408	0.7178 ± 0.0738	0.7722 ± 0.0655	0.5600 ± 0.0870	0.8303 ± 0.0570
	ResNet50	0.7148 ± 0.0620	0.7244 ± 0.0656	0.4314 ± 0.1249	0.7302 ± 0.0590	0.7752 ± 0.0532	0.6233 ± 0.1029	0.8062 ± 0.0323
	VGG16	0.6501 ± 0.0503	0.6640 ± 0.0511	0.3064 ± 0.0962	0.6772 ± 0.0455	0.7359 ± 0.0495	0.5322 ± 0.1501	0.7679 ± 0.1094
Miv vs Mis	DenseNet121	0.5919 ± 0.0261	0.7075 ± 0.0395	0.2077 ± 0.0498	0.7074 ± 0.0328	0.7325 ± 0.0262	0.8976 ± 0.0167	0.2862 ± 0.0671
	ResNet50	0.6429 ± 0.0545	0.7315 ± 0.0335	0.2905 ± 0.0811	0.7442 ± 0.0427	0.7492 ± 0.0236	0.8527 ± 0.0705	0.4331 ± 0.1716
	VGG16	0.5485 ± 0.0599	0.6733 ± 0.0660	0.1144 ± 0.1330	0.6555 ± 0.1033	0.7015 ± 0.0416	0.9711 ± 0.0356	0.1259 ± 0.1554
Multiclass	DenseNet121	0.5336 ± 0.0298	0.5494 ± 0.0412	0.4419 ± 0.0289	0.5648 ± 0.0405	0.7202 ± 0.0214	-	-
	ResNet50	0.5057 ± 0.0331	0.5176 ± 0.0552	0.3679 ± 0.0403	0.5456 ± 0.0552	0.6987 ± 0.0375	-	-
	VGG16	0.4734 ± 0.0465	0.4911 ± 0.0568	0.3256 ± 0.0997	0.5124 ± 0.0605	0.6744 ± 0.0475	-	-



(a) BT < 0.8 vs ≥ 0.8 mm.

(b) Miv vs Mis.

(c) Multiclass.

Fig. 1: Radar plots with the average results across the 5-fold cross-validation for each metric, model and task. Lef: Breslow thickness classification task (BT < 0.8 vs ≥ 0.8 mm). Center: melanoma in situ vs. invasive (Miv vs Mis). Right: multiclass classification task (Mis vs Miv with BT < 0.8 vs Miv with ≥ 0.8 mm).

II. Performance per fold

TABLE II: **Results obtained for the supervised and semi-supervised learning approaches on the BT classification task.** The performance of the models is evaluated using different metrics, which are reported for each of the cross-validation folds, as well as for their average (together with their standard deviation).

		Balanced acc	**F1-score**	**Kappa score**	**Precision**	**AUC**	**Recall**	**Specificity**
Supervised learning	Fold 1	0.7080	0.7439	0.4175	0.7435	0.7912	0.6027	0.8133
	Fold 2	0.6507	0.6382	0.2972	0.6648	0.7275	0.5106	0.7907
	Fold 3	0.7688	0.7714	0.5363	0.7718	0.8254	0.7500	0.7876
	Fold 4	0.6458	0.6590	0.2984	0.6608	0.6991	0.5176	0.7739
	Fold 5	0.8006	0.8094	0.6078	0.8101	0.8329	0.7356	0.8655
	Average	*0.7148 ± 0.0620*	*0.7244 ± 0.0656*	*0.4314 ± 0.1249*	*0.7302 ± 0.0590*	*0.7752 ± 0.0532*	*0.6233 ± 0.1029*	*0.8062 ± 0.0323*
Semi-supervised learning	Fold 1	0.7897	0.8123	0.5753	0.8131	0.8714	0.7260	0.8533
	Fold 2	0.7334	0.7242	0.4611	0.7499	0.8274	0.6064	0.8605
	Fold 3	0.7619	0.7660	0.5244	0.7659	0.8453	0.7273	0.7965
	Fold 4	0.5977	0.6011	0.2120	0.6637	0.6948	0.2824	0.9130
	Fold 5	0.7922	0.8001	0.5889	0.8001	0.8768	0.7356	0.8487
	Average	*0.7350 ± 0.0719*	*0.7407 ± 0.0762*	*0.4724 ± 0.1377*	*0.7585 ± 0.0526*	*0.8232 ± 0.0666*	*0.6155 ± 0.1733*	*0.8544 ± 0.0371*

TABLE III: **Results obtained for the supervised and semi-supervised learning approaches on the Mis versus Miv classification task.** The performance of the models is evaluated using different metrics, which are reported for each of the cross-validation folds as well as for their average (together with their standard deviation).

		Balanced acc	**F1-score**	**Kappa score**	**Precision**	**AUC**	**Recall**	**Specificity**
Supervised learning	Fold 1	0.6204	0.6865	0.2487	0.6827	0.7171	0.8033	0.4375
	Fold 2	0.6698	0.7745	0.3332	0.7768	0.7708	0.8483	0.4912
	Fold 3	0.6426	0.7056	0.3027	0.7023	0.7253	0.8421	0.4430
	Fold 4	0.7234	0.7638	0.4045	0.7827	0.7743	0.7842	0.6625
	Fold 5	0.5582	0.7268	0.1634	0.7766	0.7584	0.9853	0.1311
	Average	*0.6429 ± 0.0545*	*0.7315 ± 0.0335*	*0.2905 ± 0.0811*	*0.7442 ± 0.0427*	*0.7492 ± 0.0236*	*0.8527 ± 0.0705*	*0.4331 ± 0.1716*
Semi-supervised learning	Fold 1	0.6090	0.6971	0.2548	0.7091	0.7864	0.9180	0.3000
	Fold 2	0.6095	0.7720	0.2669	0.7696	0.8014	0.9384	0.2807
	Fold 3	0.6650	0.7364	0.3689	0.7455	0.8076	0.9123	0.4177
	Fold 4	0.7127	0.8089	0.4683	0.8085	0.8547	0.9253	0.5000
	Fold 5	0.6739	0.7941	0.3904	0.7915	0.7923	0.9216	0.4262
	Average	*0.6540 ± 0.0399*	*0.7617 ± 0.0405*	*0.3499 ± 0.0800*	*0.7648 ± 0.0350*	*0.8085 ± 0.0242*	*0.9231 ± 0.0088*	*0.3849 ± 0.0826*

TABLE IV: **Results obtained for the supervised and semi-supervised learning approaches on the multiclass classification task (Mis versus Miv with BT $<$ 0.8 mm versus Miv with BT $\geq$ 0.8 mm.** The performance of the models is evaluated using different metrics, which are reported for each of the cross-validation folds as well as for their average (together with their standard deviation).

		Balanced acc	**F1-score**	**Kappa score**	**Precision**	**AUC**
Supervised learning	Fold 1	0.4891	0.4842	0.4154	0.5031	0.7044
	Fold 2	0.5497	0.5841	0.4473	0.5815	0.7087
	Fold 3	0.5208	0.5259	0.4441	0.5320	0.6961
	Fold 4	0.5299	0.5541	0.4917	0.6057	0.7493
	Fold 5	0.5785	0.5988	0.4109	0.6016	0.7423
	Average	*0.5336 ± 0.0298*	*0.5494 ± 0.0412*	*0.4419 ± 0.0289*	*0.5648 ± 0.0405*	*0.7202 ± 0.0214*
Semi-supervised learning	Fold 1	0.5765	0.5751	0.5062	0.6006	0.7785
	Fold 2	0.6233	0.6554	0.5937	0.6600	0.7787
	Fold 3	0.5188	0.5040	0.3869	0.5261	0.7089
	Fold 4	0.5591	0.5848	0.5177	0.6061	0.7738
	Fold 5	0.5709	0.6065	0.4227	0.6152	0.7455
	Average	*0.5697 ± 0.0335*	*0.5852 ± 0.0491*	*0.4854 ± 0.0733*	*0.6016 ± 0.0432*	*0.7571 ± 0.0271*

III. Gradient maps

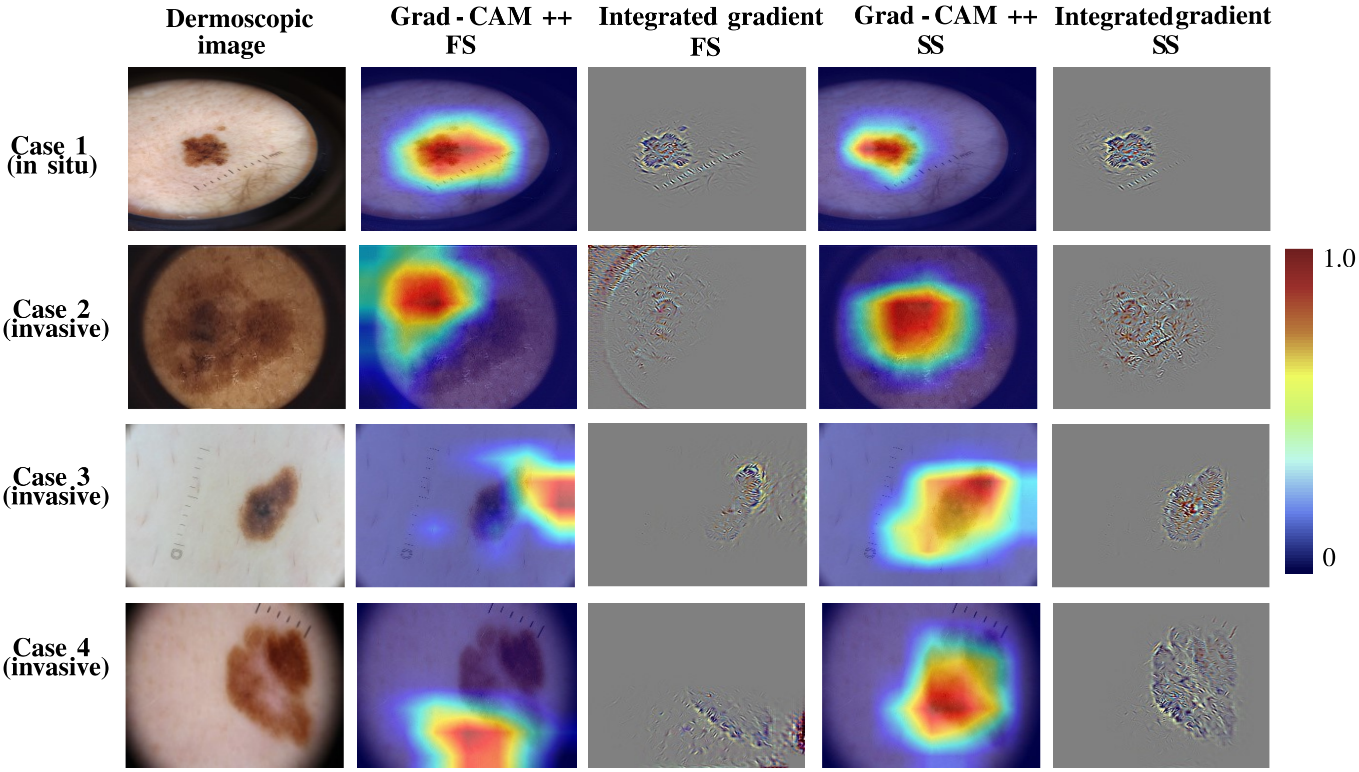


Fig. 2: **Grad-CAM++ and integrated gradient plot.** This image matrix shows four cases for the Mis vs. Miv classification task where the semi-supervised model outperforms the supervised one. The first image on the left corresponds to the image of the melanocytic lesion, followed by Grad-CAM++ with its corresponding integrated gradient for the supervised and semi-supervised models, respectively. In all cases, semi-supervised models show better delimitation of the lesions, a fact that is also verified by the corresponding integrated gradients. FS, fully supervised; SS, semisupervised.