

Table S1: Class-wise comparison of dermoscopy-only and DDCF models: three-partition summaries for each backbone

Backbone	Lesion	Support per partition	Recall	Specificity	F1	AUROC
Swin-Tiny	AKIEC	61	0.4973 $\pm$ 0.1617 / 0.4536 $\pm$ 0.0683	0.9362 $\pm$ 0.0369 / 0.9676 $\pm$ 0.0267	0.3923 $\pm$ 0.0226 / 0.4691 $\pm$ 0.0692	0.8866 $\pm$ 0.0126 / 0.9008 $\pm$ 0.0184
	BCC	504	0.7705 $\pm$ 0.0850 / 0.8651 $\pm$ 0.0865	0.9039 $\pm$ 0.0360 / 0.8783 $\pm$ 0.0313	0.8215 $\pm$ 0.0408 / 0.8662 $\pm$ 0.0391	0.9322 $\pm$ 0.0196 / 0.9464 $\pm$ 0.0124
	BKL	109	0.3884 $\pm$ 0.1452 / 0.5107 $\pm$ 0.0852	0.9426 $\pm$ 0.0251 / 0.9447 $\pm$ 0.0172	0.4031 $\pm$ 0.0838 / 0.5143 $\pm$ 0.0477	0.8018 $\pm$ 0.0388 / 0.8439 $\pm$ 0.0360
	DF	11	0.4848 $\pm$ 0.1389 / 0.3333 $\pm$ 0.2288	0.9828 $\pm$ 0.0248 / 0.9968 $\pm$ 0.0020	0.4403 $\pm$ 0.2726 / 0.3972 $\pm$ 0.2708	0.8570 $\pm$ 0.0463 / 0.9667 $\pm$ 0.0225
	INF	10	0.6333 $\pm$ 0.1528 / 0.3333 $\pm$ 0.2082	0.9760 $\pm$ 0.0224 / 0.9932 $\pm$ 0.0051	0.3856 $\pm$ 0.2082 / 0.3470 $\pm$ 0.2293	0.9133 $\pm$ 0.0639 / 0.9265 $\pm$ 0.0373
	MEL	90	0.5630 $\pm$ 0.0932 / 0.5444 $\pm$ 0.0778	0.9560 $\pm$ 0.0146 / 0.9627 $\pm$ 0.0153	0.5537 $\pm$ 0.0293 / 0.5612 $\pm$ 0.0113	0.9243 $\pm$ 0.0189 / 0.9333 $\pm$ 0.0058
	NV	149	0.6890 $\pm$ 0.1040 / 0.7450 $\pm$ 0.0293	0.9614 $\pm$ 0.0120 / 0.9528 $\pm$ 0.0019	0.7153 $\pm$ 0.0445 / 0.7349 $\pm$ 0.0163	0.9388 $\pm$ 0.0156 / 0.9480 $\pm$ 0.0040
	SCC/KA	95	0.6912 $\pm$ 0.0438 / 0.6737 $\pm$ 0.0316	0.9470 $\pm$ 0.0129 / 0.9562 $\pm$ 0.0153	0.6243 $\pm$ 0.0404 / 0.6408 $\pm$ 0.0488	0.9311 $\pm$ 0.0055 / 0.9361 $\pm$ 0.0173
	VASC	9	0.6667 $\pm$ 0.1111 / 0.7037 $\pm$ 0.1283	0.9951 $\pm$ 0.0029 / 0.9958 $\pm$ 0.0015	0.6028 $\pm$ 0.0210 / 0.6429 $\pm$ 0.0908	0.9901 $\pm$ 0.0066 / 0.9759 $\pm$ 0.0368
DenseNet121	AKIEC	61	0.4153 $\pm$ 0.0095 / 0.4426 $\pm$ 0.0656	0.9116 $\pm$ 0.0006 / 0.9597 $\pm$ 0.0059	0.2934 $\pm$ 0.0047 / 0.4227 $\pm$ 0.0348	0.8167 $\pm$ 0.0190 / 0.8920 $\pm$ 0.0149
	BCC	504	0.6131 $\pm$ 0.0503 / 0.7811 $\pm$ 0.0961	0.9132 $\pm$ 0.0273 / 0.8677 $\pm$ 0.0421	0.7184 $\pm$ 0.0285 / 0.8110 $\pm$ 0.0431	0.9017 $\pm$ 0.0215 / 0.9219 $\pm$ 0.0049
	BKL	109	0.3089 $\pm$ 0.0347 / 0.4557 $\pm$ 0.0972	0.9254 $\pm$ 0.0232 / 0.9211 $\pm$ 0.0475	0.3179 $\pm$ 0.0102 / 0.4298 $\pm$ 0.0127	0.7147 $\pm$ 0.0230 / 0.8090 $\pm$ 0.0098
	DF	11	0.5455 $\pm$ 0.0000 / 0.3333 $\pm$ 0.1892	0.9630 $\pm$ 0.0219 / 0.9968 $\pm$ 0.0011	0.2466 $\pm$ 0.1070 / 0.3939 $\pm$ 0.1689	0.8460 $\pm$ 0.0332 / 0.8990 $\pm$ 0.0418
	INF	10	0.3667 $\pm$ 0.1155 / 0.2333 $\pm$ 0.0577	0.9799 $\pm$ 0.0083 / 0.9922 $\pm$ 0.0042	0.2202 $\pm$ 0.0696 / 0.2294 $\pm$ 0.0066	0.9347 $\pm$ 0.0173 / 0.9255 $\pm$ 0.0514
	MEL	90	0.6000 $\pm$ 0.0962 / 0.5556 $\pm$ 0.0111	0.9473 $\pm$ 0.0092 / 0.9599 $\pm$ 0.0080	0.5545 $\pm$ 0.0379 / 0.5625 $\pm$ 0.0280	0.9113 $\pm$ 0.0198 / 0.9048 $\pm$ 0.0272
	NV	149	0.6555 $\pm$ 0.0571 / 0.7047 $\pm$ 0.0533	0.9666 $\pm$ 0.0057 / 0.9561 $\pm$ 0.0097	0.7061 $\pm$ 0.0390 / 0.7162 $\pm$ 0.0143	0.9461 $\pm$ 0.0070 / 0.9421 $\pm$ 0.0108
	SCC/KA	95	0.6632 $\pm$ 0.0365 / 0.6877 $\pm$ 0.1054	0.9021 $\pm$ 0.0209 / 0.9346 $\pm$ 0.0072	0.5047 $\pm$ 0.0312 / 0.5862 $\pm$ 0.0496	0.9043 $\pm$ 0.0078 / 0.9163 $\pm$ 0.0131
	VASC	9	0.7778 $\pm$ 0.1925 / 0.7037 $\pm$ 0.0642	0.9942 $\pm$ 0.0019 / 0.9984 $\pm$ 0.0015	0.6318 $\pm$ 0.0604 / 0.7481 $\pm$ 0.0714	0.9785 $\pm$ 0.0179 / 0.9784 $\pm$ 0.0105
EfficientNet-B4	AKIEC	61	0.4918 $\pm$ 0.0591 / 0.3224 $\pm$ 0.0379	0.9120 $\pm$ 0.0558 / 0.9672 $\pm$ 0.0106	0.3571 $\pm$ 0.0853 / 0.3516 $\pm$ 0.0593	0.8206 $\pm$ 0.0364 / 0.8598 $\pm$ 0.0174
	BCC	504	0.6726 $\pm$ 0.1253 / 0.8049 $\pm$ 0.0648	0.8652 $\pm$ 0.0468 / 0.8333 $\pm$ 0.0216	0.7365 $\pm$ 0.0715 / 0.8116 $\pm$ 0.0302	0.8889 $\pm$ 0.0132 / 0.9044 $\pm$ 0.0033
	BKL	109	0.2783 $\pm$ 0.0597 / 0.4434 $\pm$ 0.0371	0.9340 $\pm$ 0.0135 / 0.9085 $\pm$ 0.0172	0.3008 $\pm$ 0.0470 / 0.3990 $\pm$ 0.0141	0.7274 $\pm$ 0.0253 / 0.7932 $\pm$ 0.0037
	DF	11	0.5758 $\pm$ 0.2624 / 0.2727 $\pm$ 0.0909	0.9753 $\pm$ 0.0141 / 0.9977 $\pm$ 0.0020	0.3023 $\pm$ 0.1392 / 0.3739 $\pm$ 0.1501	0.9438 $\pm$ 0.0222 / 0.9121 $\pm$ 0.0397
	INF	10	0.2000 $\pm$ 0.1000 / 0.2333 $\pm$ 0.0577	0.9861 $\pm$ 0.0115 / 0.9919 $\pm$ 0.0020	0.1525 $\pm$ 0.0448 / 0.2245 $\pm$ 0.0353	0.8770 $\pm$ 0.0381 / 0.9435 $\pm$ 0.0174
	MEL	90	0.5148 $\pm$ 0.0421 / 0.5259 $\pm$ 0.1701	0.9501 $\pm$ 0.0166 / 0.9466 $\pm$ 0.0306	0.5056 $\pm$ 0.0266 / 0.4976 $\pm$ 0.0455	0.9012 $\pm$ 0.0160 / 0.9106 $\pm$ 0.0088
	NV	149	0.6465 $\pm$ 0.0983 / 0.6130 $\pm$ 0.1706	0.9524 $\pm$ 0.0172 / 0.9606 $\pm$ 0.0157	0.6679 $\pm$ 0.0498 / 0.6540 $\pm$ 0.0995	0.9250 $\pm$ 0.0221 / 0.9393 $\pm$ 0.0196
	SCC/KA	95	0.6246 $\pm$ 0.0580 / 0.6351 $\pm$ 0.0438	0.9325 $\pm$ 0.0161 / 0.9502 $\pm$ 0.0073	0.5444 $\pm$ 0.0130 / 0.5961 $\pm$ 0.0103	0.8917 $\pm$ 0.0157 / 0.9196 $\pm$ 0.0067
	VASC	9	0.7778 $\pm$ 0.2222 / 0.7037 $\pm$ 0.0642	0.9913 $\pm$ 0.0093 / 0.9958 $\pm$ 0.0015	0.5873 $\pm$ 0.1132 / 0.6449 $\pm$ 0.0190	0.9860 $\pm$ 0.0163 / 0.9931 $\pm$ 0.0033

Values are mean $\pm$ standard deviation across three stratified lesion-level data partitions generated using three different dataset-splitting random seeds, shown as dermoscopy-only / DDCF. Support is the number of test lesions per class in each partition; exact class-wise counts were held fixed while lesion identities differed. Support is not a pooled count, and the metrics are not from a single partition. Recall corresponds to class-wise sensitivity. Results for DF, INF, and VASC should be interpreted cautiously because of small support.

Table S2: Selected Swin-Tiny prediction transitions on the original first-experiment test partition

Outcome	True class	Dermoscopy-only	DDCF	Count
Corrected	BCC	SCC/KA	BCC	16
Corrected	MEL	NV	MEL	10
Corrected	BCC	BKL	BCC	8
Corrected	SCC/KA	BKL	SCC/KA	6
Corrected	BKL	SCC/KA	BKL	3
Adversely changed	SCC/KA	SCC/KA	BKL	3
Adversely changed	BKL	BKL	SCC/KA	2

Corrected transitions indicate cases in which DDCF recovered the true diagnosis after an incorrect dermoscopy-only prediction. The adversely changed transitions shown here correspond to the BKL–SCC/KA pair specifically raised in the review. Counts are from the original test partition used in the first Swin-Tiny experiment, which also yielded the 103 corrected and 60 adversely changed cases.

Swin-Tiny representative corrected and adversely changed cases

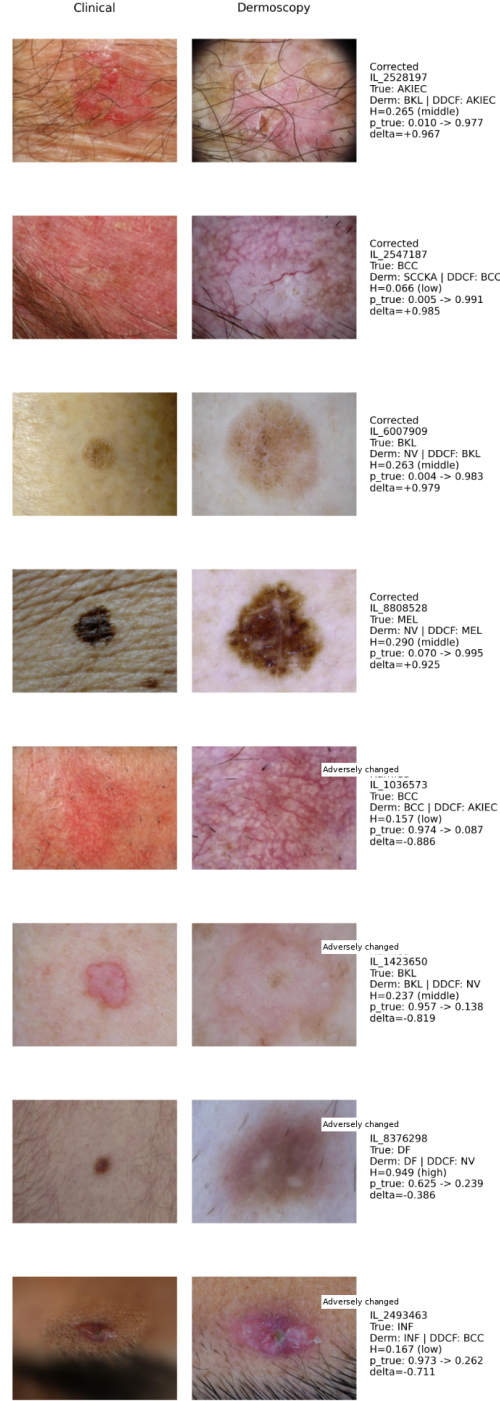


Figure S1: **Representative Swin-Tiny cases corrected or adversely changed by DDCF.** Paired clinical and dermoscopic images are shown with the true diagnosis, dermoscopy-only prediction, DDCF prediction, dermoscopy-only predictive entropy, and change in the true-class probability. Corrected cases were misclassified by the dermoscopy-only model but correctly classified by DDCF, whereas adversely changed cases were correctly classified by dermoscopy alone but misclassified by DDCF. Clinical features such as photodamage, elevation, erythema, and surrounding skin changes were not formally annotated and are therefore presented for qualitative interpretation only. Examples are from the original first-experiment Swin-Tiny test partition used for the case-level analysis in Table S2.

Swin-Tiny low/high derm-only uncertainty: AKIEC

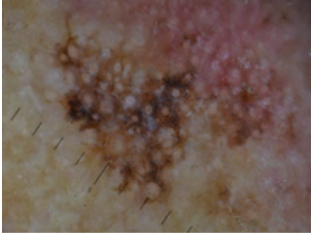
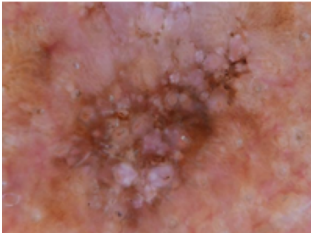
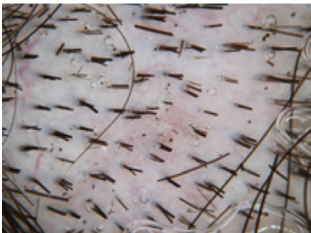
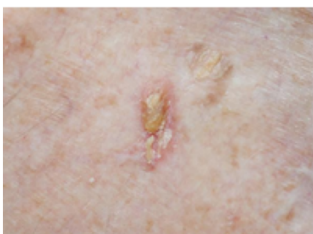
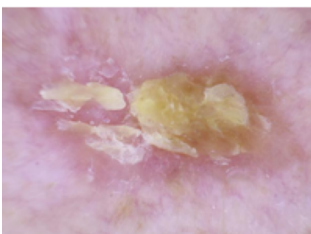
Clinical	Dermoscopy	
		Lowest entropy IL_1418667 True: AKIEC Derm: AKIEC DDCF: AKIEC H=0.032 (low) p_true: 0.996 -> 1.000 delta=+0.004 Outcome: both_correct
		Lowest entropy IL_8987309 True: AKIEC Derm: AKIEC DDCF: AKIEC H=0.053 (low) p_true: 0.993 -> 0.998 delta=+0.005 Outcome: both_correct
		Highest entropy IL_1085478 True: AKIEC Derm: BKL DDCF: NV H=1.572 (high) p_true: 0.031 -> 0.064 delta=+0.033 Outcome: both_wrong
		Highest entropy IL_1122729 True: AKIEC Derm: INF DDCF: AKIEC H=1.375 (high) p_true: 0.067 -> 0.549 delta=+0.483 Outcome: corrected

Figure S2: **Representative low- and high-uncertainty cases across diagnostic categories.** For each diagnostic category, two cases with the lowest and two cases with the highest predictive entropy from the Swin-Tiny dermoscopy-only model are presented. Paired clinical and dermoscopic images are shown together with the true diagnosis, dermoscopy-only prediction, DDCF prediction, predictive entropy, true-class probability change, and paired prediction outcome. Examples are from the original first-experiment Swin-Tiny test partition.

Swin-Tiny low/high derm-only uncertainty: BCC

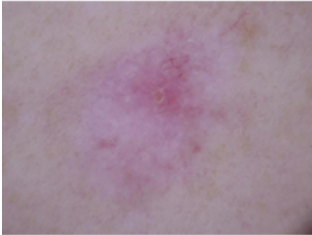
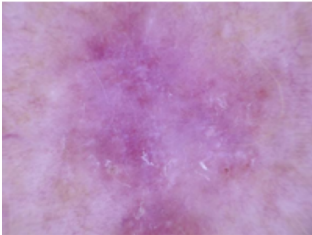
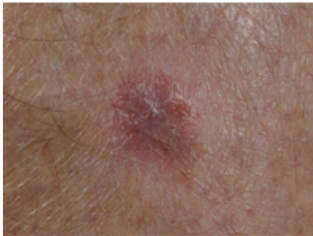
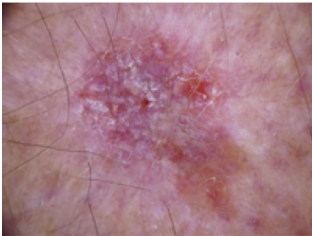
Clinical	Dermoscopy	
		Lowest entropy IL_7965758 True: BCC Derm: INF DDCF: INF H=0.047 (low) p_true: 0.004 -> 0.446 delta=+0.442 Outcome: both_wrong
		Lowest entropy IL_4782455 True: BCC Derm: INF DDCF: INF H=0.056 (low) p_true: 0.003 -> 0.008 delta=+0.004 Outcome: both_wrong
		Highest entropy IL_3888766 True: BCC Derm: SCCKA DDCF: AKIEC H=1.624 (high) p_true: 0.217 -> 0.014 delta=-0.203 Outcome: both_wrong
		Highest entropy IL_0784671 True: BCC Derm: AKIEC DDCF: BCC H=1.508 (high) p_true: 0.202 -> 0.630 delta=+0.428 Outcome: corrected

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: BKL

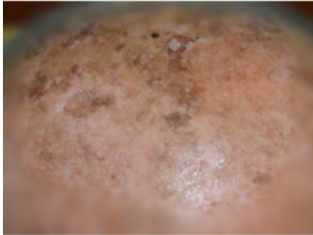
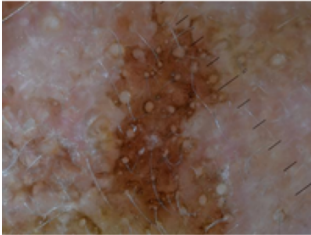
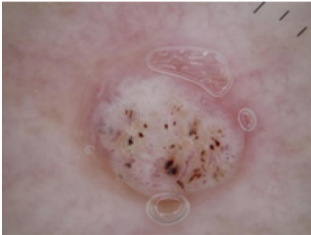
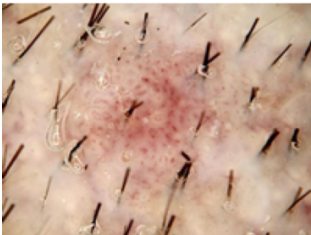
Clinical	Dermoscopy	
		Lowest entropy IL_6361071 True: BKL Derm: BKL DDCF: BKL $H=0.052$ (low) $p_{\text{true}}: 0.993 \rightarrow 0.981$ $\text{delta}=-0.011$ Outcome: both_correct
		Lowest entropy IL_2514915 True: BKL Derm: BKL DDCF: BKL $H=0.055$ (low) $p_{\text{true}}: 0.993 \rightarrow 0.798$ $\text{delta}=-0.194$ Outcome: both_correct
		Highest entropy IL_1187197 True: BKL Derm: BCC DDCF: BCC $H=1.587$ (high) $p_{\text{true}}: 0.317 \rightarrow 0.182$ $\text{delta}=-0.135$ Outcome: both_wrong
		Highest entropy IL_3052885 True: BKL Derm: BCC DDCF: BKL $H=1.455$ (high) $p_{\text{true}}: 0.236 \rightarrow 0.736$ $\text{delta}=+0.499$ Outcome: corrected

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: DF

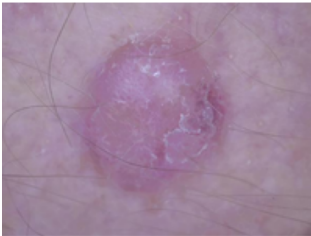
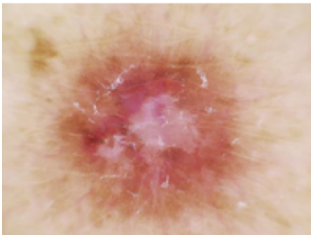
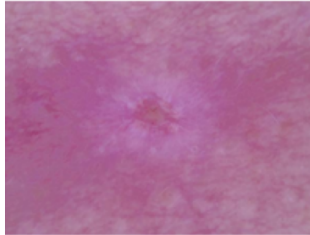
Clinical	Dermoscopy	
		Lowest entropy IL_3884905 True: DF Derm: DF DDCF: DF H=0.103 (low) p_true: 0.985 -> 0.976 delta=-0.009 Outcome: both_correct
		Lowest entropy IL_2128198 True: DF Derm: INF DDCF: INF H=0.115 (low) p_true: 0.001 -> 0.023 delta=+0.022 Outcome: both_wrong
		Highest entropy IL_3853188 True: DF Derm: INF DDCF: INF H=1.748 (high) p_true: 0.083 -> 0.004 delta=-0.079 Outcome: both_wrong
		Highest entropy IL_3861295 True: DF Derm: DF DDCF: DF H=1.505 (high) p_true: 0.399 -> 0.482 delta=+0.082 Outcome: both_correct

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: INF

Clinical

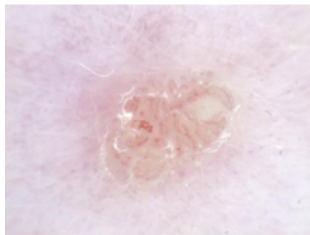
Dermoscopy



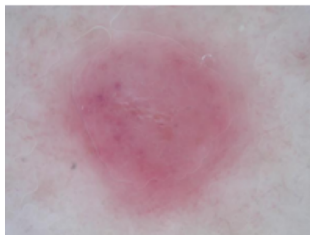
Lowest entropy
IL_6660320
True: INF
Derm: INF | DDCF: INF
H=0.037 (low)
p_true: 0.995 -> 0.939
delta=-0.057
Outcome: both_correct



Lowest entropy
IL_0485034
True: INF
Derm: DF | DDCF: DF
H=0.108 (low)
p_true: 0.002 -> 0.000
delta=-0.002
Outcome: both_wrong



Highest entropy
IL_9439830
True: INF
Derm: BKL | DDCF: NV
H=1.243 (high)
p_true: 0.040 -> 0.151
delta=+0.111
Outcome: both_wrong



Highest entropy
IL_4566979
True: INF
Derm: INF | DDCF: VASC
H=1.079 (high)
p_true: 0.568 -> 0.133
delta=-0.435
Outcome: harmed

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: MEL

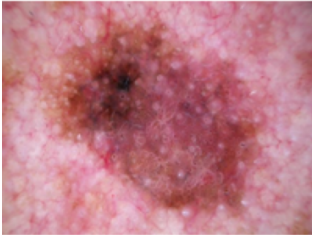
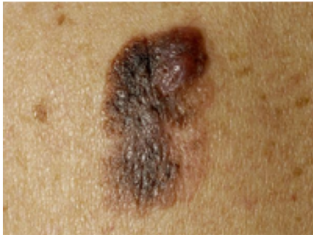
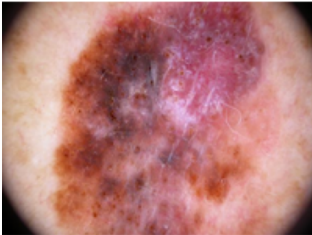
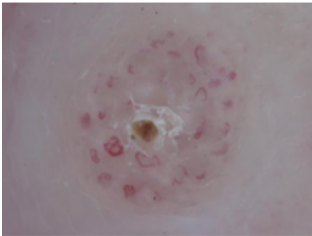
Clinical	Dermoscopy	
		Lowest entropy IL_3770255 True: MEL Derm: MEL DDCF: MEL H=0.021 (low) p_true: 0.998 -> 0.984 delta=-0.014 Outcome: both_correct
		Lowest entropy IL_1027638 True: MEL Derm: MEL DDCF: MEL H=0.027 (low) p_true: 0.997 -> 0.999 delta=+0.002 Outcome: both_correct
		Highest entropy IL_9191697 True: MEL Derm: BKL DDCF: VASC H=1.803 (high) p_true: 0.009 -> 0.008 delta=-0.001 Outcome: both_wrong
		Highest entropy IL_1275816 True: MEL Derm: NV DDCF: NV H=1.589 (high) p_true: 0.121 -> 0.285 delta=+0.164 Outcome: both_wrong

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: NV

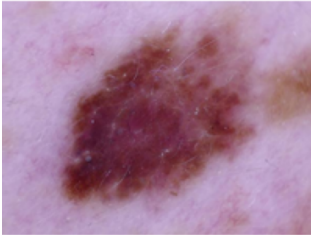
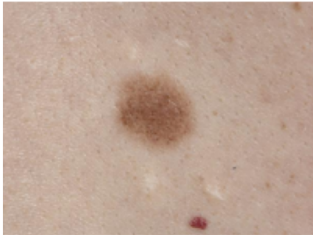
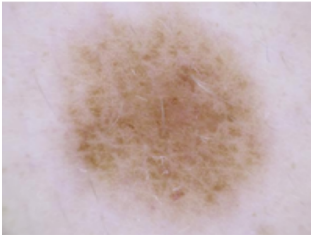
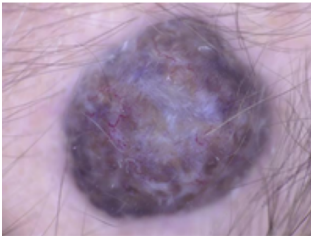
Clinical	Dermoscopy	
		Lowest entropy IL_4106945 True: NV Derm: MEL DDCF: NV H=0.048 (low) p_true: 0.005 -> 0.780 delta=+0.775 Outcome: corrected
		Lowest entropy IL_4072310 True: NV Derm: NV DDCF: NV H=0.064 (low) p_true: 0.990 -> 0.999 delta=+0.009 Outcome: both_correct
		Highest entropy IL_0047878 True: NV Derm: MEL DDCF: MEL H=1.454 (high) p_true: 0.236 -> 0.135 delta=-0.101 Outcome: both_wrong
		Highest entropy IL_6889294 True: NV Derm: NV DDCF: NV H=1.314 (high) p_true: 0.422 -> 0.546 delta=+0.124 Outcome: both_correct

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: SCCKA

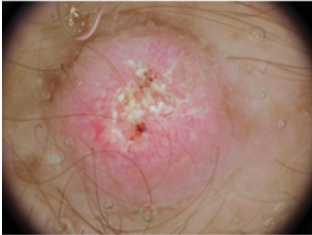
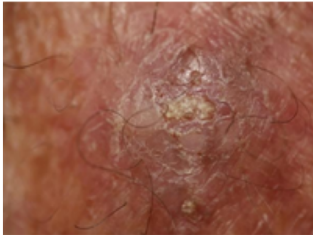
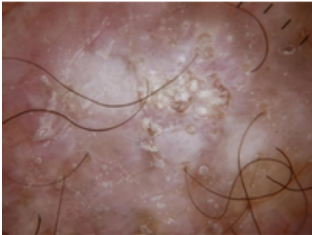
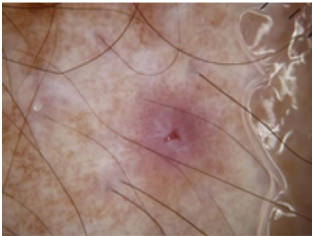
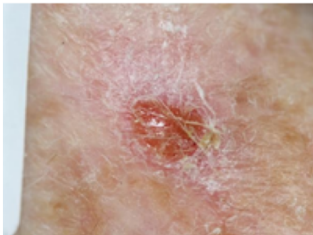
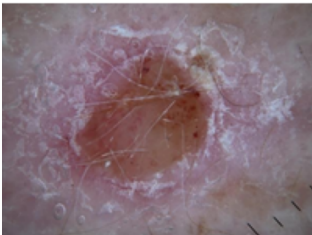
Clinical	Dermoscopy	
		Lowest entropy IL_4614364 True: SCCKA Derm: SCCKA DDCF: SCCKA H=0.026 (low) p_true: 0.997 -> 0.999 delta=+0.003 Outcome: both_correct
		Lowest entropy IL_5083460 True: SCCKA Derm: SCCKA DDCF: SCCKA H=0.031 (low) p_true: 0.996 -> 1.000 delta=+0.004 Outcome: both_correct
		Highest entropy IL_8726676 True: SCCKA Derm: BCC DDCF: AKIEC H=1.754 (high) p_true: 0.240 -> 0.087 delta=-0.153 Outcome: both_wrong
		Highest entropy IL_8813361 True: SCCKA Derm: BCC DDCF: SCCKA H=1.373 (high) p_true: 0.280 -> 0.890 delta=+0.610 Outcome: corrected

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Swin-Tiny low/high derm-only uncertainty: VASC

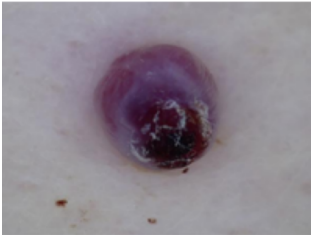
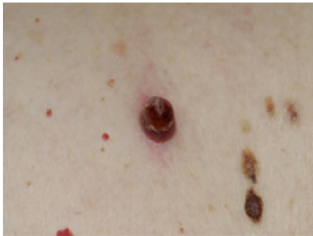
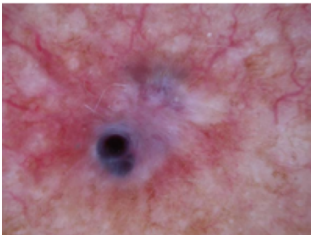
Clinical	Dermoscopy	
		Lowest entropy IL_9783995 True: VASC Derm: VASC DDCF: VASC $H=0.017$ (low) $p_{\text{true}}: 0.998 \rightarrow 0.999$ $\text{delta}=+0.001$ Outcome: both_correct
		Lowest entropy IL_7687935 True: VASC Derm: VASC DDCF: VASC $H=0.025$ (low) $p_{\text{true}}: 0.997 \rightarrow 0.999$ $\text{delta}=+0.002$ Outcome: both_correct
		Highest entropy IL_6019983 True: VASC Derm: VASC DDCF: VASC $H=1.378$ (high) $p_{\text{true}}: 0.365 \rightarrow 0.990$ $\text{delta}=+0.625$ Outcome: both_correct
		Highest entropy IL_3067885 True: VASC Derm: BCC DDCF: VASC $H=1.320$ (high) $p_{\text{true}}: 0.212 \rightarrow 0.781$ $\text{delta}=+0.569$ Outcome: corrected

Figure S2: Representative low- and high-uncertainty cases across diagnostic categories (continued).

Table S3: Paired DDCF–naive-sum differences within each backbone and data partition

Backbone	Partition	Δ macro-AUROC (95% CI)	Δ macro-F1 (95% CI)	Δ BACC (95% CI)
Swin-Tiny	P1	+0.0148 (−0.0020, +0.0311)	+0.0417 (−0.0070, +0.0874)	−0.0585 (−0.1134, −0.0052)
Swin-Tiny	P2	+0.0134 (−0.0028, +0.0311)	+0.0692 (+0.0152, +0.1150)	−0.0529 (−0.1065, −0.0029)
Swin-Tiny	P3	+0.0045 (−0.0081, +0.0183)	+0.0403 (−0.0115, +0.0907)	−0.0363 (−0.0910, +0.0154)
DenseNet121	P1	+0.0008 (−0.0113, +0.0124)	+0.0405 (−0.0094, +0.0836)	−0.0161 (−0.0636, +0.0292)
DenseNet121	P2	+0.0306 (+0.0102, +0.0497)	+0.0911 (+0.0431, +0.1326)	+0.0082 (−0.0373, +0.0501)
DenseNet121	P3	+0.0044 (−0.0102, +0.0182)	+0.0112 (−0.0362, +0.0585)	+0.0172 (−0.0368, +0.0684)
EfficientNet-B4	P1	+0.0098 (−0.0034, +0.0238)	+0.0075 (−0.0479, +0.0634)	+0.0177 (−0.0384, +0.0762)
EfficientNet-B4	P2	+0.0136 (+0.0027, +0.0244)	+0.0635 (+0.0082, +0.1148)	−0.1022 (−0.1636, −0.0419)
EfficientNet-B4	P3	+0.0103 (−0.0017, +0.0229)	+0.0305 (−0.0339, +0.0912)	+0.0033 (−0.0664, +0.0731)

Differences are DDCF ($\lambda = 0.7$) minus naive-sum. Each test partition contained 1,038 lesions. Point differences were calculated on the original test partition; intervals are the 2.5th and 97.5th percentiles from 10,000 paired class-stratified bootstrap replicates. Each replicate preserved class support and used identical lesion indices for both methods. The analysis was performed separately for each backbone and partition; overlapping partitions were not pooled. P1–P3 identify evaluations within each backbone and do not imply three independent cohorts or a verified common partition mapping across backbone families. Intervals are nominal, unadjusted for multiple comparisons, and conditional on the fitted models and observed class composition. They do not provide a confidence interval for the three-partition mean or account for all training, configuration-selection, or unobserved patient-level dependence. Source predictions were paired by lesion ID with matching true labels. The original first-experiment case analysis is described separately in Table S2 and Figures S1–S2.

Supplementary benchmark implementation details

FusionM4Net was adapted from the available implementation while preserving its original overall network architecture, including the dual ResNet-50 backbone structure and additive feature-fusion design. TFormer was adapted from the official implementation while retaining its original architecture, including the Swin-Tiny encoder and cross-modal transformer fusion blocks; for the MILK10k experiments, the encoder was instantiated through a `timm`-based adaptation layer. FDDM was reimplemented for the paired MILK10k setting while preserving the original fundus-enhanced distillation architecture, including the ResNet-50 clinical-image teacher and dermoscopic-image student, knowledge-distillation pathway, and feature-alignment mechanism. CRD-Net was adapted from the available implementation while retaining its original CrossMaResNet architecture, including dual ResNet-18 encoders and cross-modal interaction modules.

Across the baselines, the original overall model architecture, backbone configuration, and principal fusion or interaction mechanism were retained. Task-specific components required for adaptation to the nine-class paired MILK10k setting, such as the data-loading interface and output classification layer, were modified where necessary. Input sizes followed the corresponding original publications and implementations. FDDM and CRD-Net used input images of 448×448 pixels.

Model-specific optimization and training hyperparameters, including the optimizer, learning rate, and batch size, followed the corresponding original publications and implementations wherever applicable, rather than a single common configuration imposed across all models. FusionM4Net and TFormer used a batch size of 32 for 50 epochs, FDDM used a batch size of 8 for up to 50 epochs, and CRD-Net used a batch size of 8 for 50 epochs. Batch sizes were fixed experimental settings; no systematic batch-size search or automatic GPU-memory maximization was performed.