Interpretations are useful: penalizing explanations to align neural networks with prior knowledge

Laura Rieger
DTU

Chandan Singh
UC Berkeley

W. James Murdoch
UC Berkeley

Bin Yu
UC Berkeley

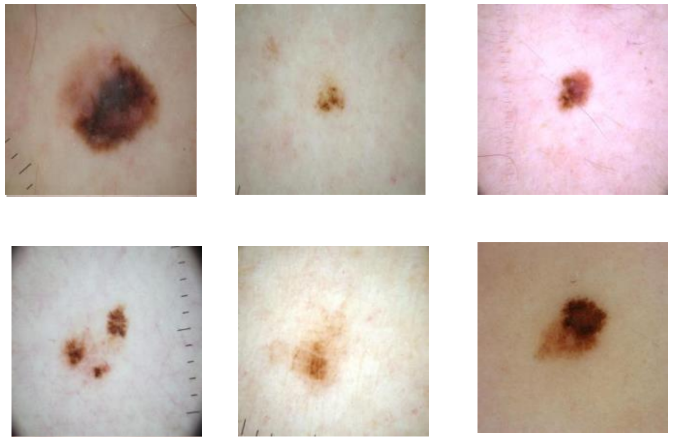
overview

datasets are biased

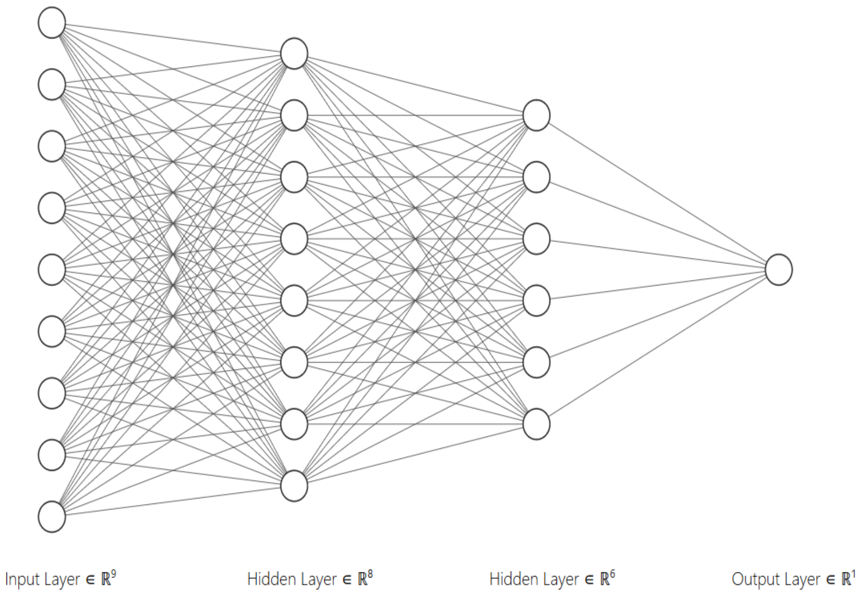
- NNs learn from large datasets
- often biased
- we sometimes know the bias

Benign

Cancerous



augmenting the loss function

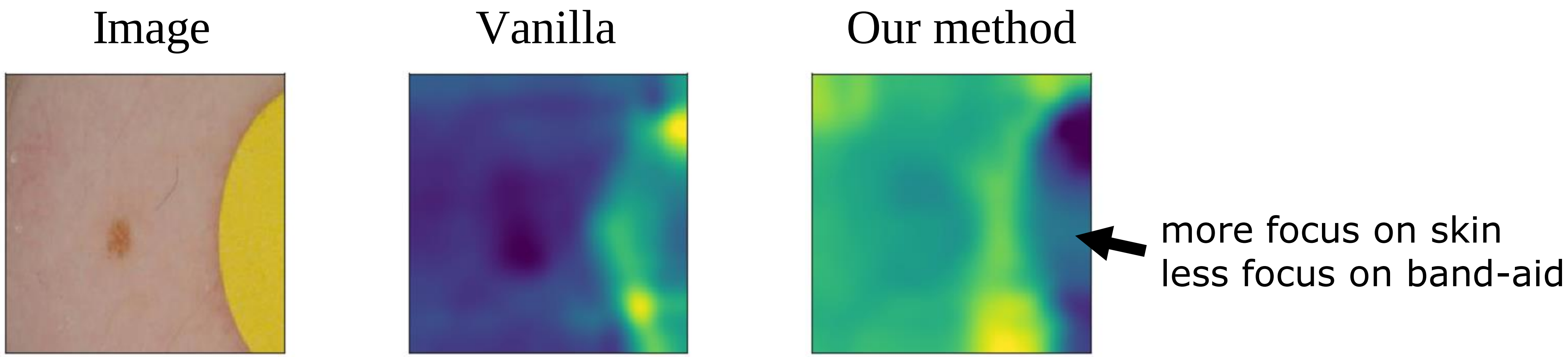


Prediction ← **True label**

Explanation ← **Prior knowledge**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \boxed{\mathcal{L}\left(f_\theta(X), y\right)} + \lambda \boxed{\mathcal{L}_{\mathrm{expl}}\left(\mathrm{expl}_\theta(X), \mathrm{expl}_X\right)}$$

using our method improves accuracy

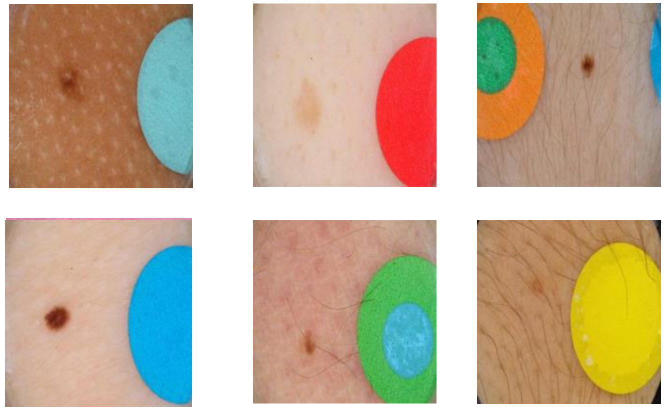


Test F1:	0.67	**0.73**

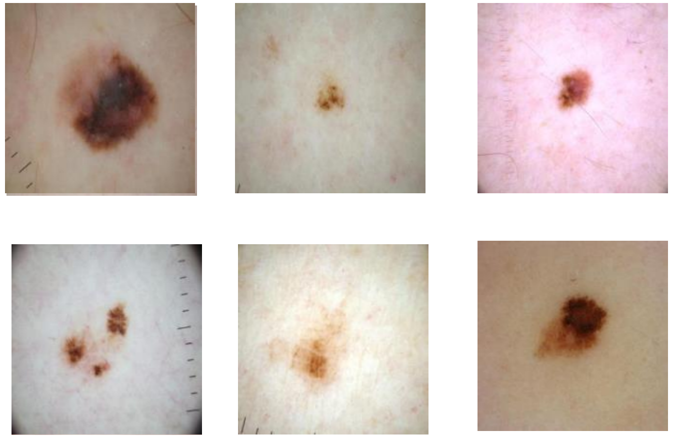
details

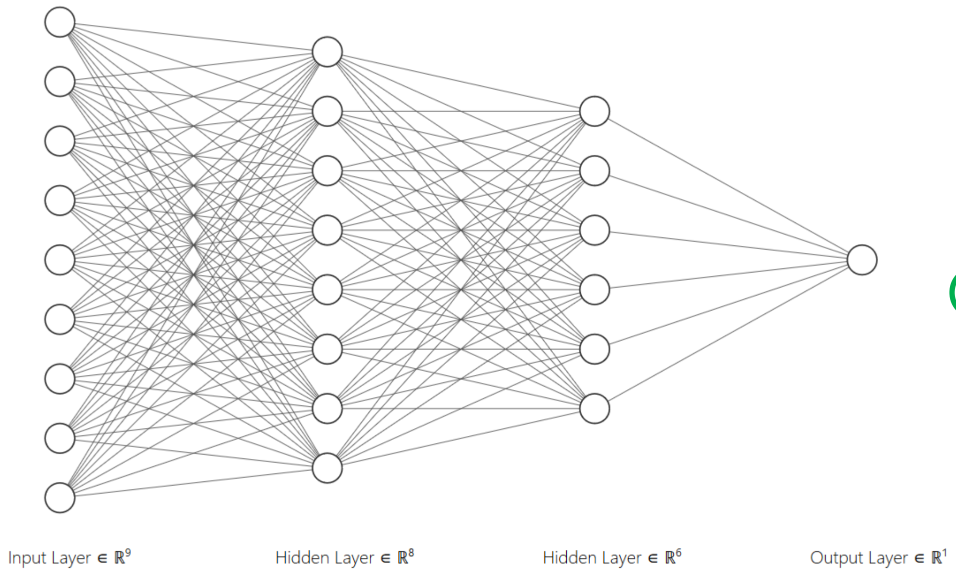
training with biased data

Benign



Cancerous

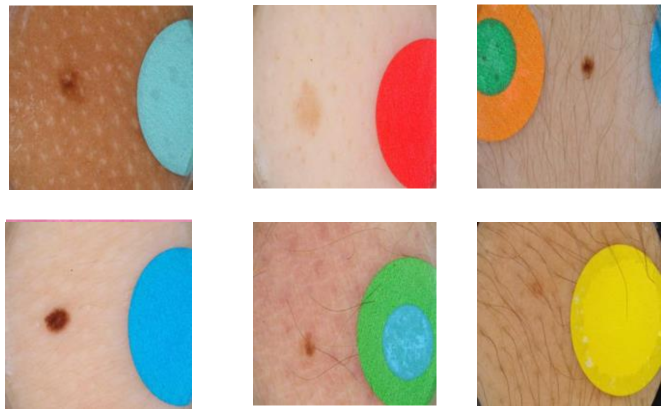




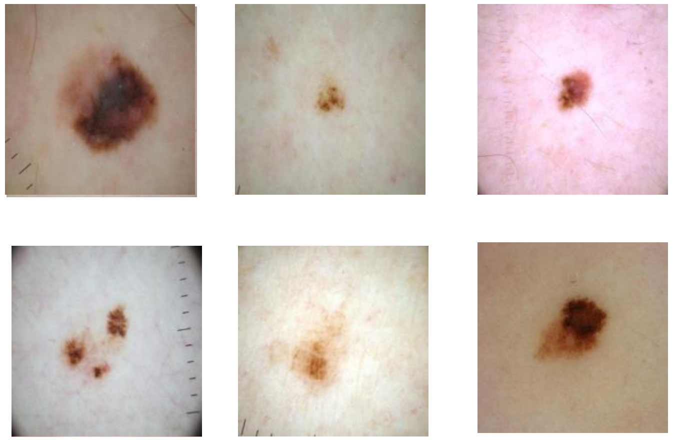
☺ 90% accurate

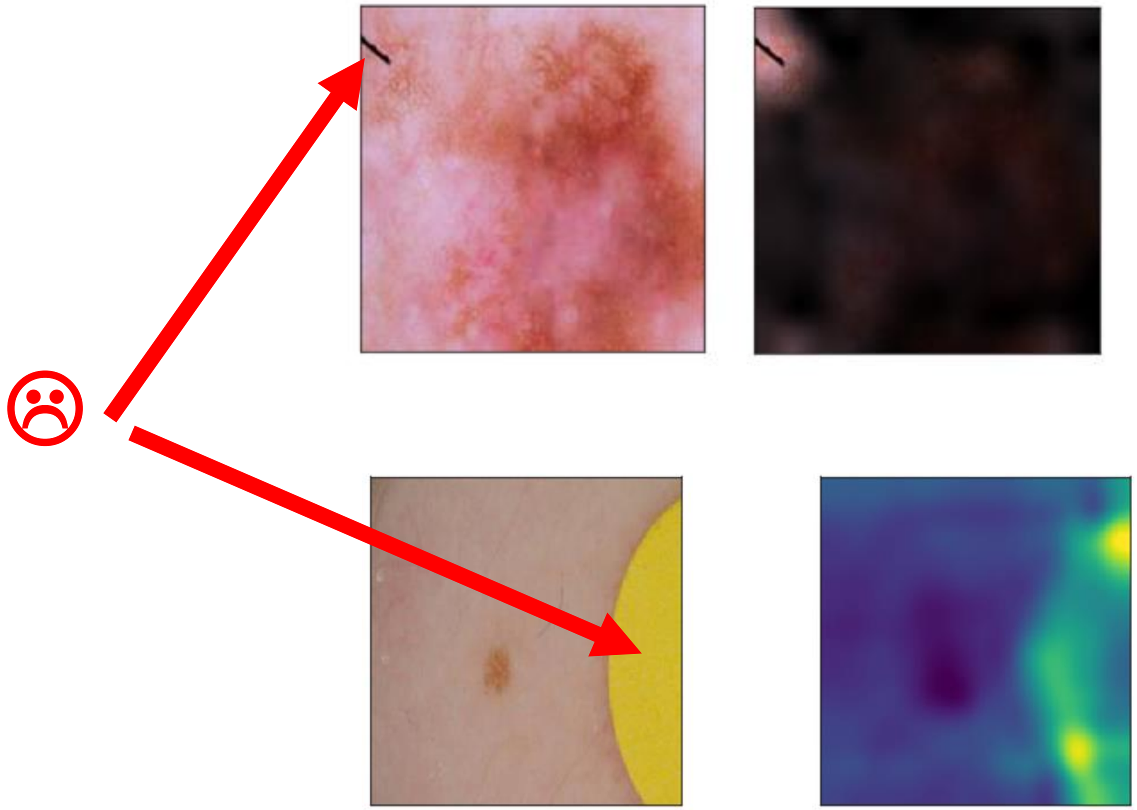
what did the network learn?

Benign



Cancerous





We know the bias (sometimes)

Gender is not important for job applications!

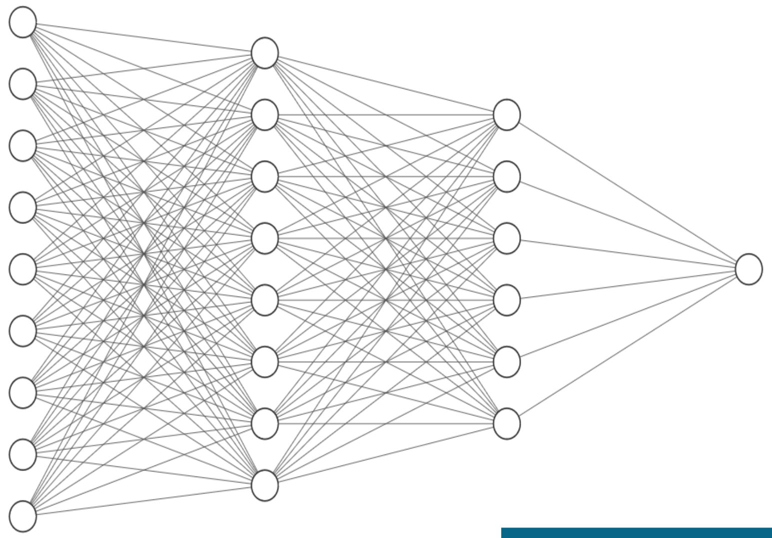
Race shouldn’t determine jail time!

Rulers aren’t cancerous!

Band aids don’t protect against cancer!

our method

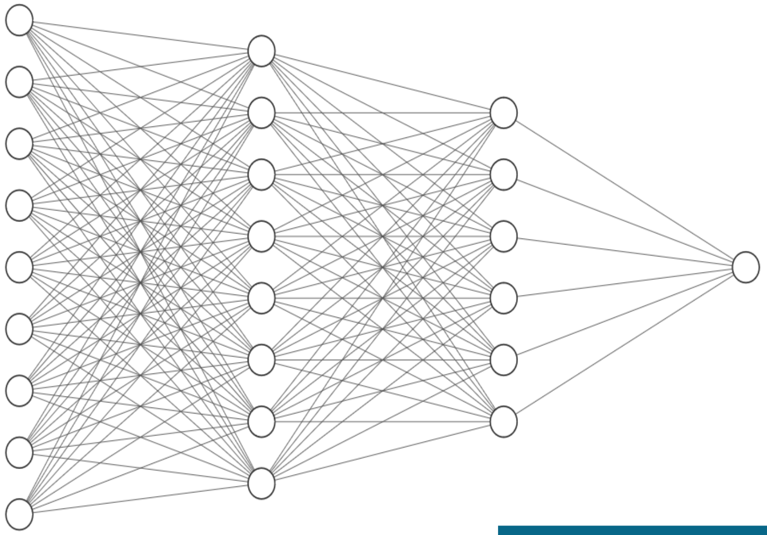
augmenting the loss function



Prediction ⟵ **True label**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \boxed{\mathcal{L}\left(f_\theta(X), y\right)}$$

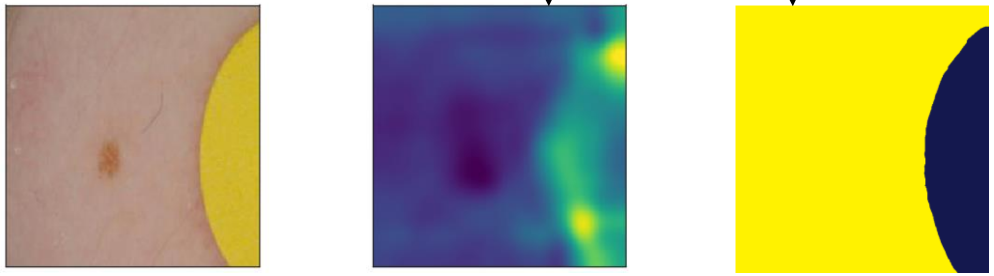
augmenting the loss function



Prediction ← **True label**

Explanation ← **Prior knowledge**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \boxed{\mathcal{L}\left(f_\theta(X), y\right)} + \lambda \boxed{\mathcal{L}_{\text{expl}}\left(\text{expl}_\theta(X), \text{expl}_X\right)}$$



Contextual **D**ecomposition **E**xplanation **P**enalty

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \mathcal{L}\left(f_\theta(X), y\right) + \lambda\, \mathcal{L}_{\text{expl}}\left(\text{expl}_\theta(X), \text{expl}_X\right)$$

any differentiable explanation method works

we used contextual decomposition (Singh 2019)

- captures interactions
- computationally lighter

[1] Singh, Chandan, W. James Murdoch, and Bin Yu. "Hierarchical interpretations for neural network predictions."

Contextual Decomposition (Singh 2019)

- requires partition of input $\{x_j\}_{j \in S}, \{x_i\}_{i \notin S}$
- iteratively forward-pass both partitions

$$g^{CD}(x) = g_L^{CD}(g_{L-1}^{CD}(...(g_2^{CD}(g_1^{CD}(x)))))$$

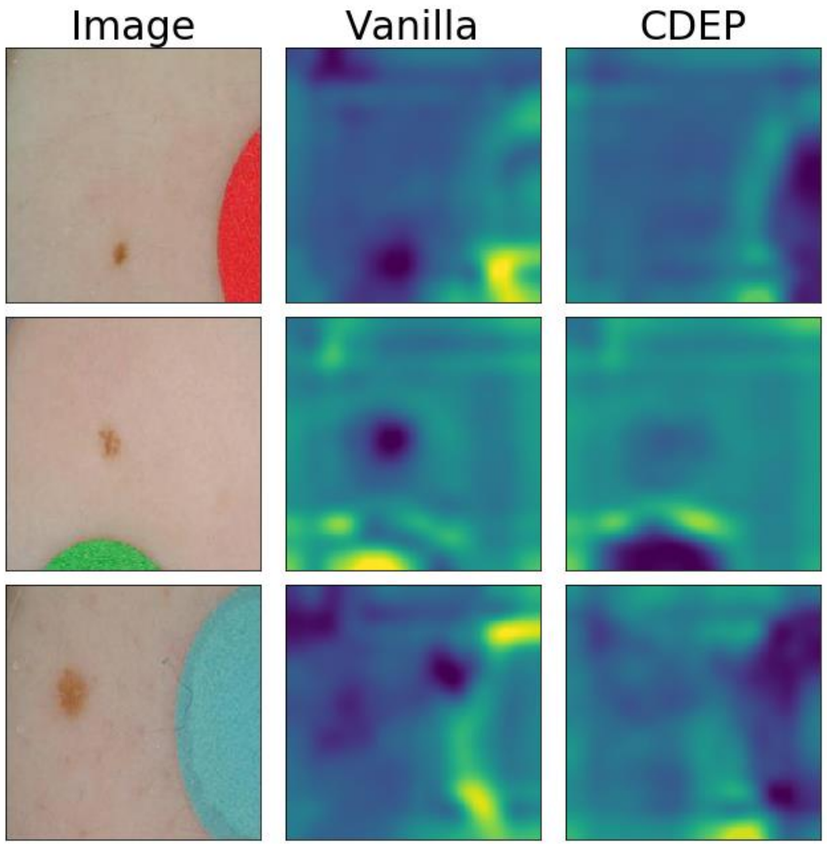
- output contribution of both partitions

$$g^{CD}(x) = (\beta(x), \gamma(x))$$

results

skin cancer (ISIC)

	AUC (NO PATCHES)	F1 (NO PATCHES)	AUC (ALL)	F1 (ALL)
VANILLA (UNBIASED DATA)	0.87	0.57	0.92	0.55
VANILLA	0.93	0.67	0.96	0.67
RRR	0.76	0.45	0.87	0.45
CDEP	**0.95**	**0.73**	**0.97**	**0.73**



explanations focus more on skin

mnist variants

	VANILLA	CDEP	RRR	EXPECTED GRADIENTS
COLORMNIST	0.2 ± 0.2	$\mathbf{31.0 \pm 2.3}$	0.2 ± 0.1	10.0 ± 0.1

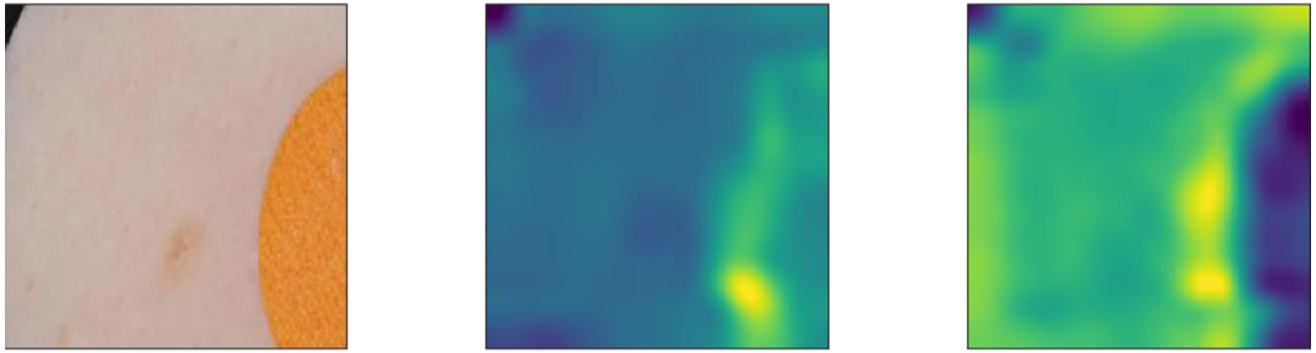
contributions

contributions

CDEP uses explainability methods to regularize an NN

used to incorporate prior knowledge into neural networks

usable with more complex knowledge than previous methods



0.67 (f1) unpenalized

0.73 (f1) penalized