Knowledge Distillation to Ensemble Global and Interpretable Prototype-based Mammogram Classification Models

Chong Wang[1], Yuanhong Chen[1], Yuyuan Liu[1], Yu Tian[1], Fengbei Liu[1], Davis J. McCarthy[2], Michael Elliott[2], Helen Frazer[3], Gustavo Carneiro[1]

[1] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, Australia
[2] St Vincent's Institute of Medical Research, Melbourne, Australia
[3] St Vincent's Hospital Melbourne, Melbourne, Australia

INTRODUCTION

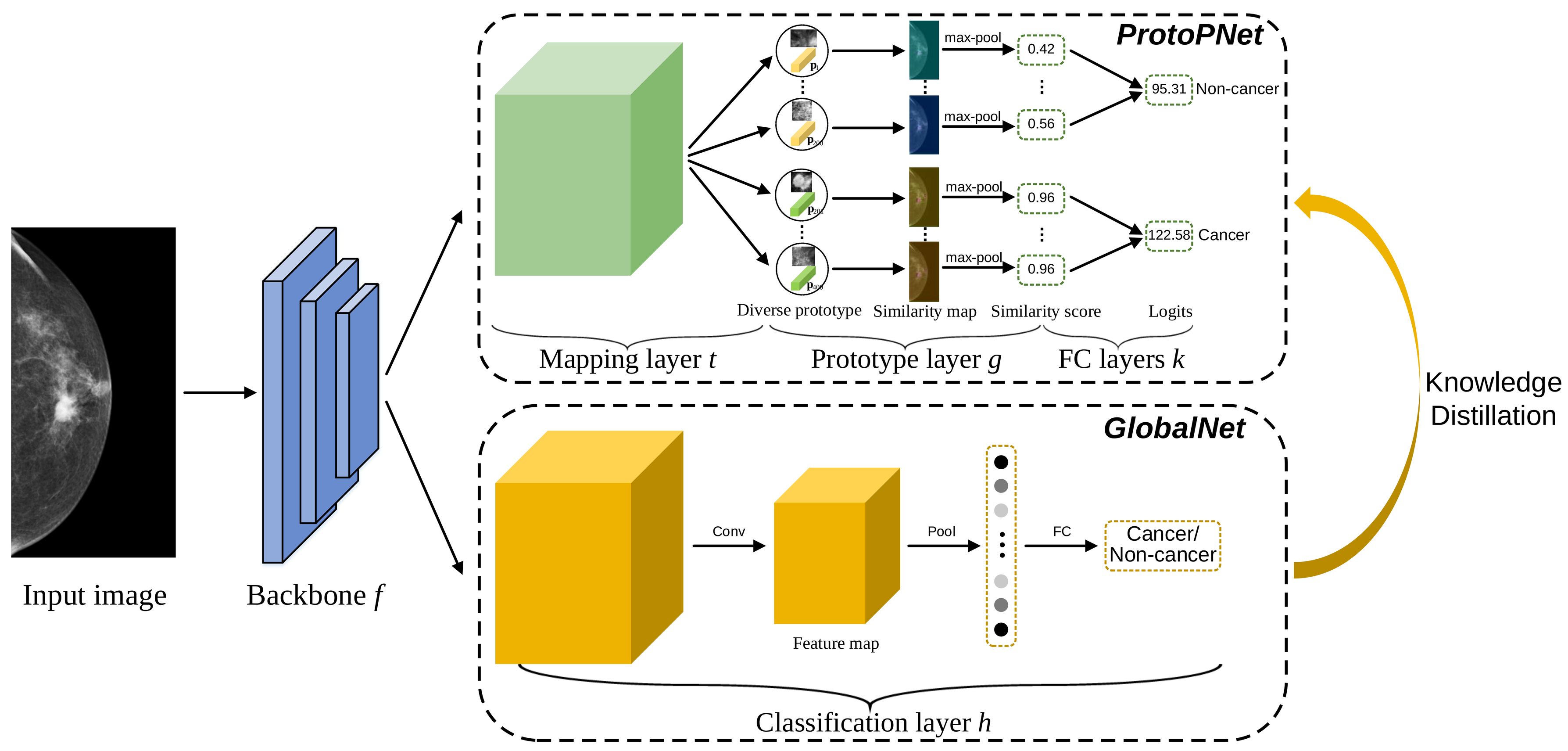
Interpretability is a key factor to the successful translation of deep-learning mammogram classifiers into real-world clinical practice.

- Prototype-based classifiers provide promising self-interpretable predictions but are less accurate than non-interpretable global whole-image classifiers
- Poor prototype diversity in prototype-based classifiers

Strategy:

- Integrate a prototype-based classifier with existing global classifiers to form a highly accurate and interpretable ensemble model
- Distill knowledge from global classifiers to improve the accuracy of the prototype-based classifier
- A greedy prototype projection strategy to improve prototype diversity

METHOD



1. Interpretable Prototype-based Classifier (ProtoPNet):

- Learn a set of class-specific prototypes $P=\{p_m\}_{m=1}^{M}$ from training samples
- Achieve interpretable classification decisions by comparing local parts of an image with training prototypes
- Cross-entropy, cluster, and separation losses to train ProtoPNet

$$l_{PPN} = l_{CE} + \lambda_1 l_{CT} + \lambda_2 l_{SP}$$

$$l_{CT} = \frac{1}{B}\sum_{i=1}^{B} \min_{p_m \in P_{y_i}} \min_{z \in Z_i} \|z - p_m\|_2^2$$

$$l_{SP} = \max\left(0, \gamma - \frac{1}{B}\sum_{i=1}^{B} \min_{p_m \notin P_{y_i}} \min_{z \in Z_i} \|z - p_m\|_2^2\right)$$

2. Knowledge Distillation:

- Enforce ProtoPNet to achieve classification accuracy as high as the non-interpretable global classifier (GlobalNet)

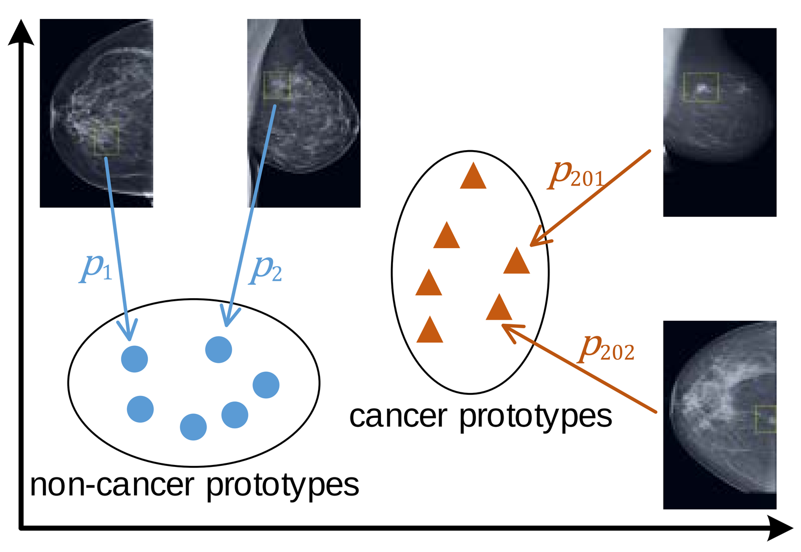
$$l_{KD} = \frac{1}{B}\sum_{i=1}^{B} \max\left(0,\ (y_i)^T(\tilde{y}_i^G) - (y_i)^T(\tilde{y}_i^P) + w\right)$$

3. Greedy Prototype Projection to Improve Prototype Diversity:

- Create an ordered prototype-image distance dictionary
- Update each prototype with the nearest unused image:

$$p_m \leftarrow \arg\min_{z \in Z} \|z - p_m\|_2^2$$

p_1	Image2, Image5, Image1, Image6, …
p_2	Image2, Image1, Image4, Image3, …
…	…
p_M	Image3, Image7, Image2, Image9, …



DATA & RESULTS

Datesets

Private ADMANI:

- BreastScreen Victoria (Australia) program from 2013 to 2019
- 20592 training images (3262 cancer images, 17330 non-cancer images)
- 2032 validation images (322 cancer images, 1710 non-cancer images)
- 22525 testing images (806 cancer images, 21719 non-cancer images)
- 410 testing images have tumour annotations to assess cancer localization

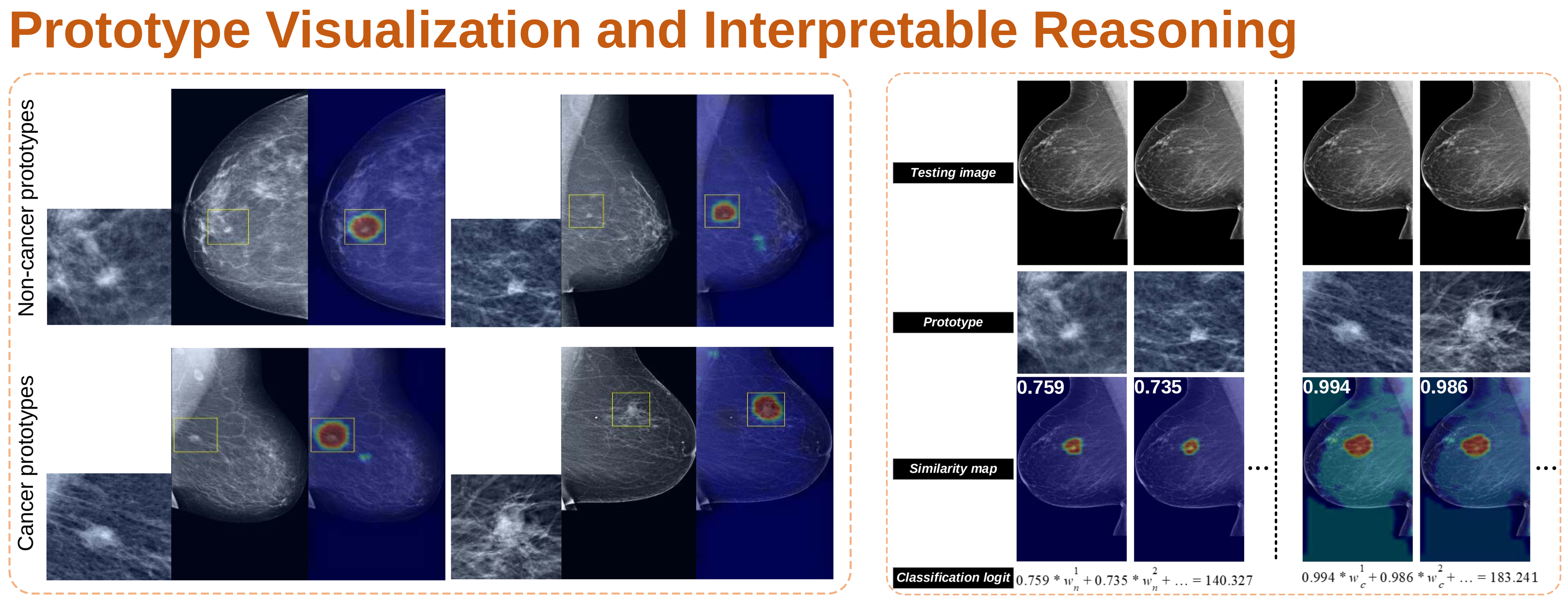
Public CMMD:

- Mammograms from 1775 Chinese patients collected from 2012 to 2016
- 2632 cancer images, 2568 non-cancer images
- The dataset is used for evaluating model's generalization ability

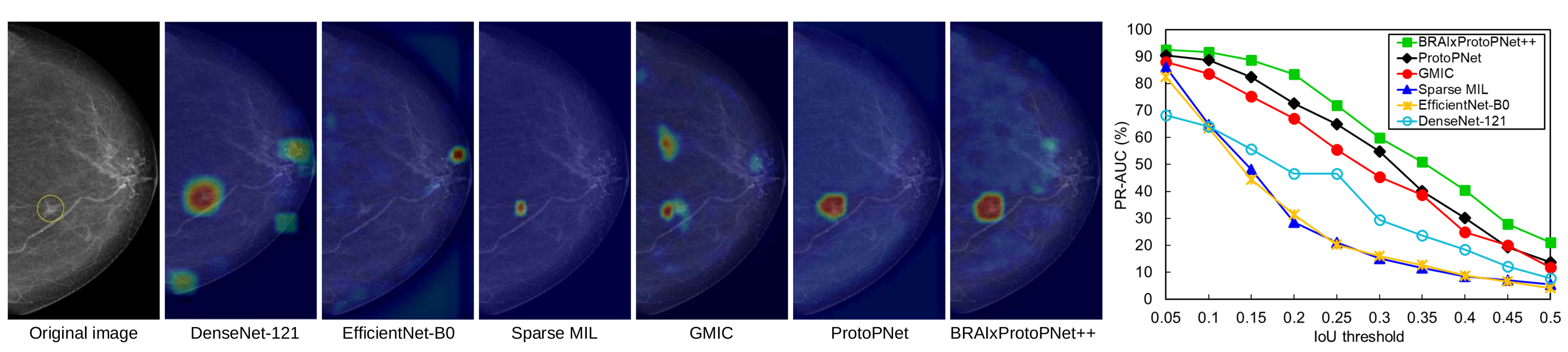
Quantitative Results

Methods			AUC ADMANI	AUC CMMD
DenseNet-121			88.54	82.38
EfficientNet-B0			89.62	76.41
Sparse MIL			89.75	81.33
GMIC			89.98	81.03
ProtoPNet (DenseNet-121)			87.12	80.23
ProtoPNet (EfficientNet-B0)			88.30	79.61
Ours (DenseNet-121)	w/o KD	ProtoPNet	87.32	80.09
		GlobalNet	88.45	82.42
		Ensemble	88.87	82.50
	w/ KD	ProtoPNet	88.35	80.67
		GlobalNet	88.61	82.52
		Ensemble	89.54	**82.65**
Ours (EfficientNet-B0)	w/o KD	ProtoPNet	88.63	79.01
		GlobalNet	90.11	76.50
		Ensemble	90.18	80.45
	w/ KD	ProtoPNet	89.55	79.86
		GlobalNet	90.12	76.47
		Ensemble	**90.68**	81.65

Prototype Visualization and Interpretable Reasoning



Breast Cancer Localization



Effect of the Greedy Prototype Projection Strategy

Methods	Cosine distance Non-cancer	Cosine distance Cancer	L2 distance Non-cancer	L2 distance Cancer	AUC
ProtoPNet w/o greedy projection	0.034	0.061	0.805	0.827	88.11
ProtoPNet w/ greedy projection	0.074	0.094	1.215	1.712	**88.30**

CONCLUSION & FUTURE PLANS

1. Prototype-based interpretability can be integrated with existing CNN classifiers to achieve interpretable and accurate mammogram classification
2. Knowledge distillation can improve the classification accuracy of the interpretable prototype-based models
3. Prototype-based interpretability can realize accurate localization results using weak image-level labels
4. Interest in applying to other medical applications, e.g., multi-class and multi-label classification

Acknowledgement

BRAIx (grant number: MRFAI000090)