DukeNUS Medical School

m.liu@u.nus.edu

Variable importance cloud for medical image analysis

Yilin Ning, PhD[1#], Mingxuan Liu, MSc[1#], Mingcheng Zhu, MRes[2], Yian Ma, MSc[2], Xin Li, MSx[2], Nan Liu, PhD[1]

[1]Duke-NUS Medical School, National University of Singapore, Singapore;
[2]Department of Engineering Science, University of Oxford, Oxford, UK

1. Background & Objective

Background

- Explainable artificial intelligence (XAI) aims to make machine learning models interpretable, especially in sensitive domains like healthcare, where trust and transparency are critical.
- Traditional XAI techniques explain only one optimal model, overlooking the variability among models with similar predictive performance.
- This narrow focus limits robustness, failing to capture uncertainty in model explanations.

Objective

- We propose a new XAI method that ensembles explanations from multiple nearly-optimal models to improve robustness and account for model uncertainty.

2. Method

❖ Datasets

- **HAM10000 skin cancer dataset[1]**
  - A large collection of 10,015 dermoscopic images of pigmented skin lesions, designed to support machine learning research in automated skin cancer diagnosis. (Outcome: skin malignancy)

❖ Model training

- Model: EfficientNet, DenseNet
- Loss: cross-entropy

Figure 1. Variable importance cloud (VIC) construction

Algorithm 1 Model Pruning and Selection

Input: optimal model M_0, validation data $\mathcal{D}_v$, original validation loss l_0, prune ratios $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, nearly-optimality ϵ, number of trail models N_0, expected number of nearly optimal models N.
Output: Selected pruned models $\mathcal{M}^*$.
$\mathcal{P}_{valid} \leftarrow \emptyset$
while $\mathcal{P}_{valid}$ is $\emptyset$ **do**
  for $p \in \mathcal{P}$ **do**
    $\mathcal{L} \leftarrow \emptyset$
    for $i = 1$ to N_0 **do**
      $l \leftarrow$ PruneModel$(M_0, p, \mathcal{D}_v)$
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{l\}$
    end for
    if $1 - \epsilon \leq \frac{\text{mean}(\mathcal{L})}{l_0} \leq 1 + \epsilon$ **then**
      $\mathcal{P}_{valid} \leftarrow \mathcal{P}_{valid} \cup \{p\}$
    end if
  end for
  if $\mathcal{P}_{valid}$ is $\emptyset$ **then**
    $\mathcal{P} = \{p'_1, p'_2, \ldots, p'_{n'}\}$ where $p'_i < p_1$
  end if
end while
Define $\mathcal{P}_{ranges} = \{(p - \delta, p + \delta) \mid p \in \mathcal{P}_{valid}\}$ where $\delta = \frac{1}{2} \min_i \{p_{i+1} - p_i\}$
for $i = 1$ to N **do**
  repeat
    Sample $p \sim U(0, 1)$
  until $p \in \cup \mathcal{P}_{ranges}$
  $M_i, l_i \leftarrow$ PruneModel$(M_0, p, \mathcal{D}_v)$
  Save M_i
end for
$\mathcal{M}^* \leftarrow \emptyset$
for $i = 1$ to N **do**
  Retrieve validation loss l_i
  if $1 - \epsilon \leq \frac{l_i}{l_0} \leq 1 + \epsilon$ **then**
    $\mathcal{M}^* \leftarrow \mathcal{M}^* \cup \{M_i\}$
  end if
end for

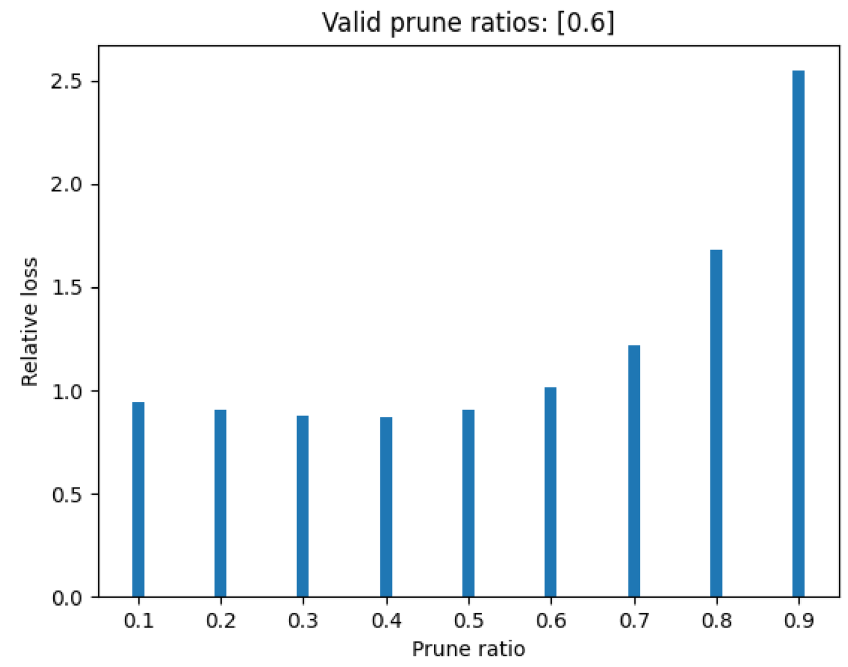
function PruneModel$(M, p, \mathcal{D}_v)$
  if log exists for p **then**
    Retrieve stored loss l
  else
    Apply pruning $M' \leftarrow$ Prune(M, p)
    Compute $l \leftarrow$ Evaluate$(M', \mathcal{D}_v)$
  end if
  return M', l
end function

❖ VIC construction

- Model pruning: Sample pruned models (i.e., masking the last convolutional layer of the model) with a random pruning ratio; retain models with mean loss within a threshold $(1 \pm \epsilon)$ to ensure near-optimality[2], as detailed in Figure 1.
- Single model explanation: Grad-CAM++
- Assembling: Aggregate Grad-CAM++ outputs across models to form a Variable Importance Cloud (VIC), capturing feature attribution variability. Both model consensus and model variability are important.

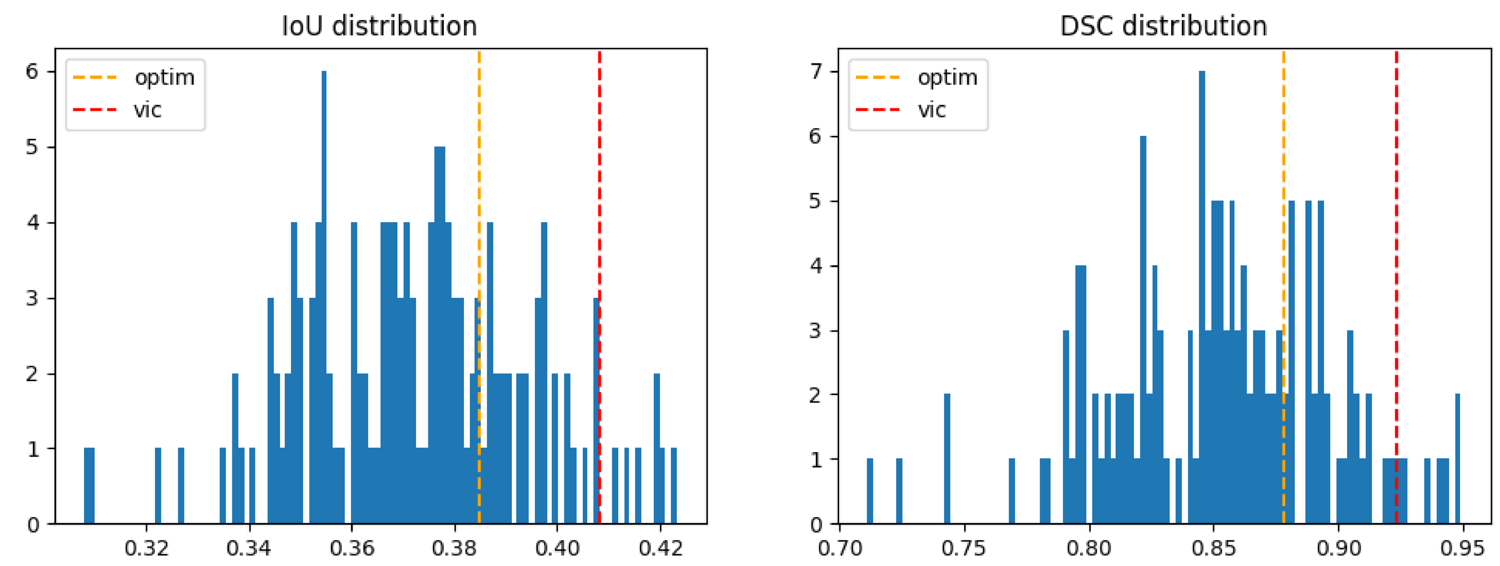
3. Result

Figure 2. Selecting valid prune ratios for nearly-optimal models



- Relative loss remains near-optimal around a pruning ratio of 0.6; higher ratios degrade performance, while lower ratios raise concerns about model similarity—making 0.6 a suitable anchor point for constructing pruned model ensembles.

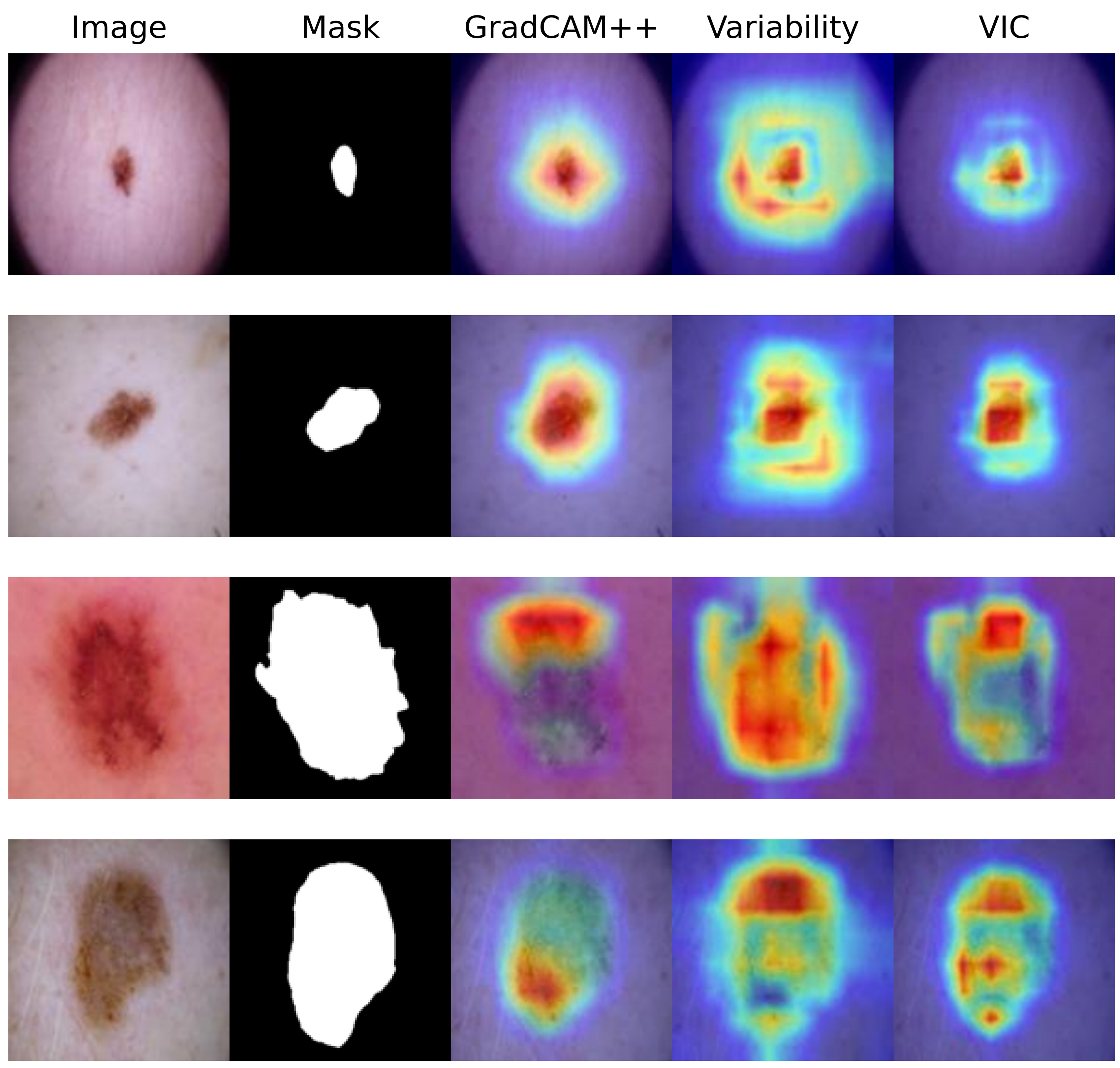
Figure 3. VIC enhances the fidelity of model explanations



	IoU	DSC
Optimal	0.385	0.878
VIC	0.408	0.923

- Nearly-optimal models display variation in fidelity of model explanations.
- Compared to the single optimal model, the VIC ensemble yields higher mean Intersection-over-Union (IoU) and Dice Similarity Coefficient (DSC), improving by 6% and 5%.

Figure 4. VIC captures more complete and stable explanations across models



- Example images show the input, ground-truth mask, Grad-CAM++ explanation from the optimal model, variability across explanations, and the VIC saliency map.
- Grad-CAM++ often highlights only part of the lesion, while VIC produces more spatially complete and coherent explanations.
- The "Variability" column reveals disagreement among nearly-optimal models, which VIC resolves into a unified and high-fidelity saliency map.

4. Conclusion & Future work

- VIC improves explanation robustness by aggregating saliency maps across nearly-optimal models, reducing reliance on a single model's potentially unstable attribution.
- Our results suggest that explanation variability is substantial, even among models with comparable performance, highlighting the need to incorporate model uncertainty into XAI frameworks.
- Future work will explore applying VIC to other explanation methods (e.g., SHAP, LIME), extend it to non-image modalities, and investigate its utility in fairness auditing and subgroup-specific interpretability.

Reference

1. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).
2. Dong J, Rudin C. Exploring the cloud of variable importance for the set of all good models. Nature Machine Intelligence 2020;2(12):810-824.