Tenth ISIC Skin Image Analysis Workshop @ MICCAI 2025

What Can We Learn from Inter-Annotator Variability in Skin Lesion Segmentation?

Kumar Abhishek†

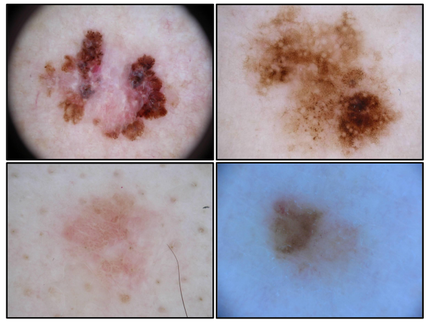
Jeremy Kawahara‡

Ghassan Hamarneh†

† SFU SIMON FRASER UNIVERSITY

‡ AIP LABS®

What Causes Variability in Medical Image Segmentation?

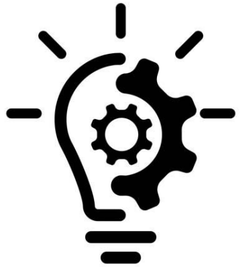


Ambiguous object boundaries

Annotators' personal preferences

Inter- and intra-annotator segmentation variability

Segmentation tools



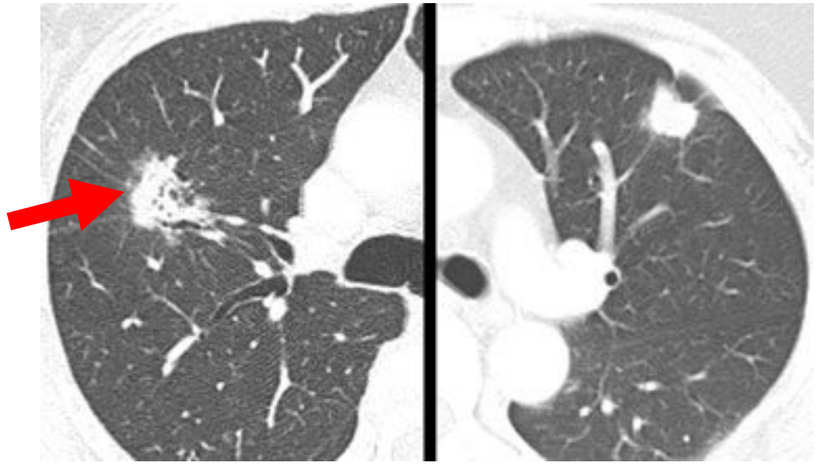
Annotators' skill levels

Lesions Without Well-Defined Boundaries

Poorly-defined boundaries are often strongly associated with **malignancy**.

Lesions Without Well-Defined Boundaries

Poorly-defined boundaries are often strongly associated with **malignancy**.



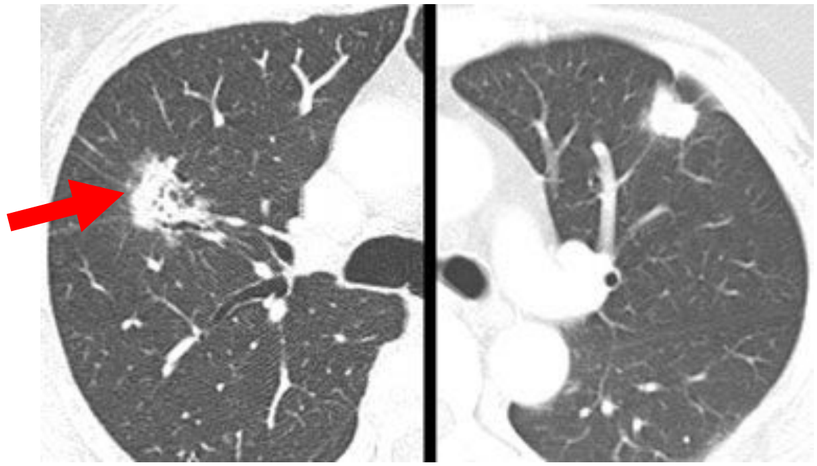
"The lesion on the **far left** has a **spicuated margin** … we should be most concerned that the lesion on the far left **is malignant**. It proved to be an **adenocarninoma** …" [1]

" … a suspicious solid **spiculated nodule** (arrow). Surgery revealed **invasive adenocarcinoma**." [2]

[1] Leung et al., 2007
[2] MacMahon et al., 2017

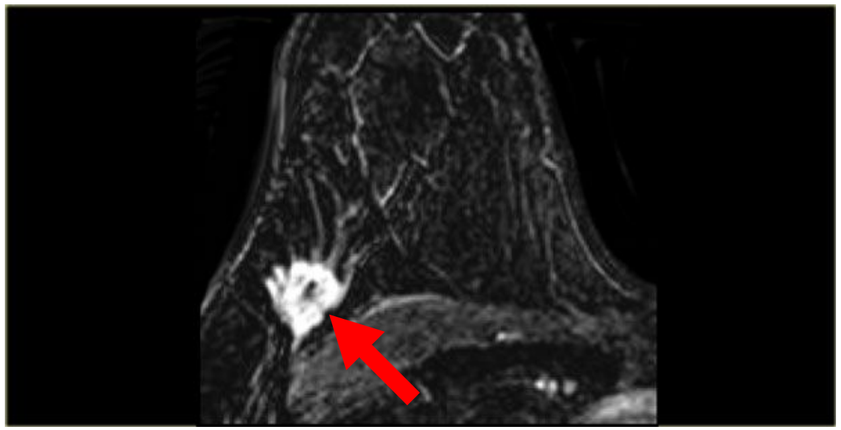
Lesions Without Well-Defined Boundaries

Poorly-defined boundaries are often strongly associated with **malignancy**.



"The lesion on the **far left** has a **spicuated margin** … we should be most concerned that the lesion on the far left **is malignant**. It proved to be an **adenocarninoma** …" [1]

" … a suspicious solid **spiculated nodule** (arrow). Surgery revealed **invasive adenocarcinoma**." [2]

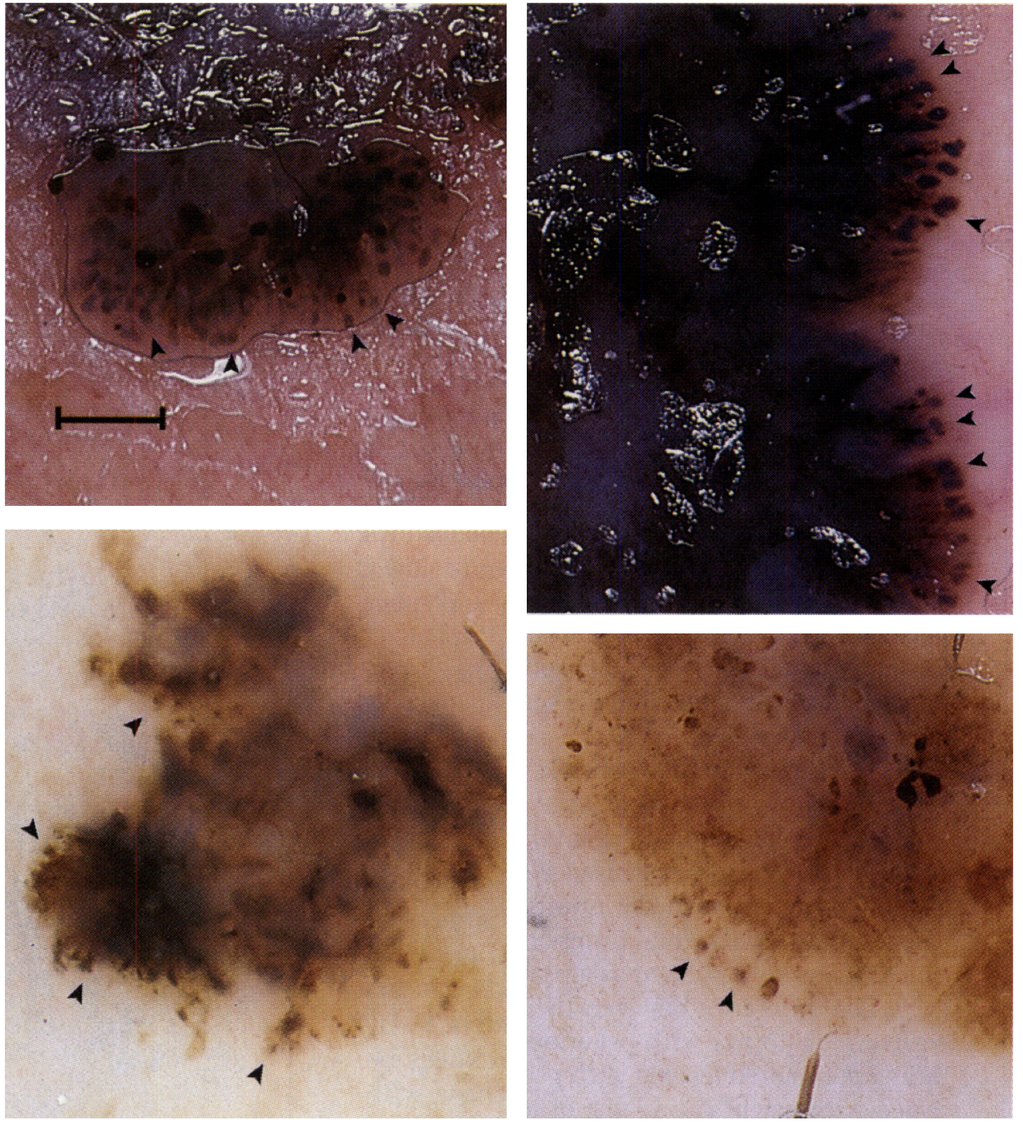


"… the image shows an irregularly shaped **mass with spiculations** and a heterogeneous internal enhancement pattern, which proved to be an **invasive lobular carcinoma**." [3]

[1] Leung et al., 2007
[2] MacMahon et al., 2017
[3] Glassman et al., 2009

Pseudopods: A Morphologic Feature in Dermoscopy

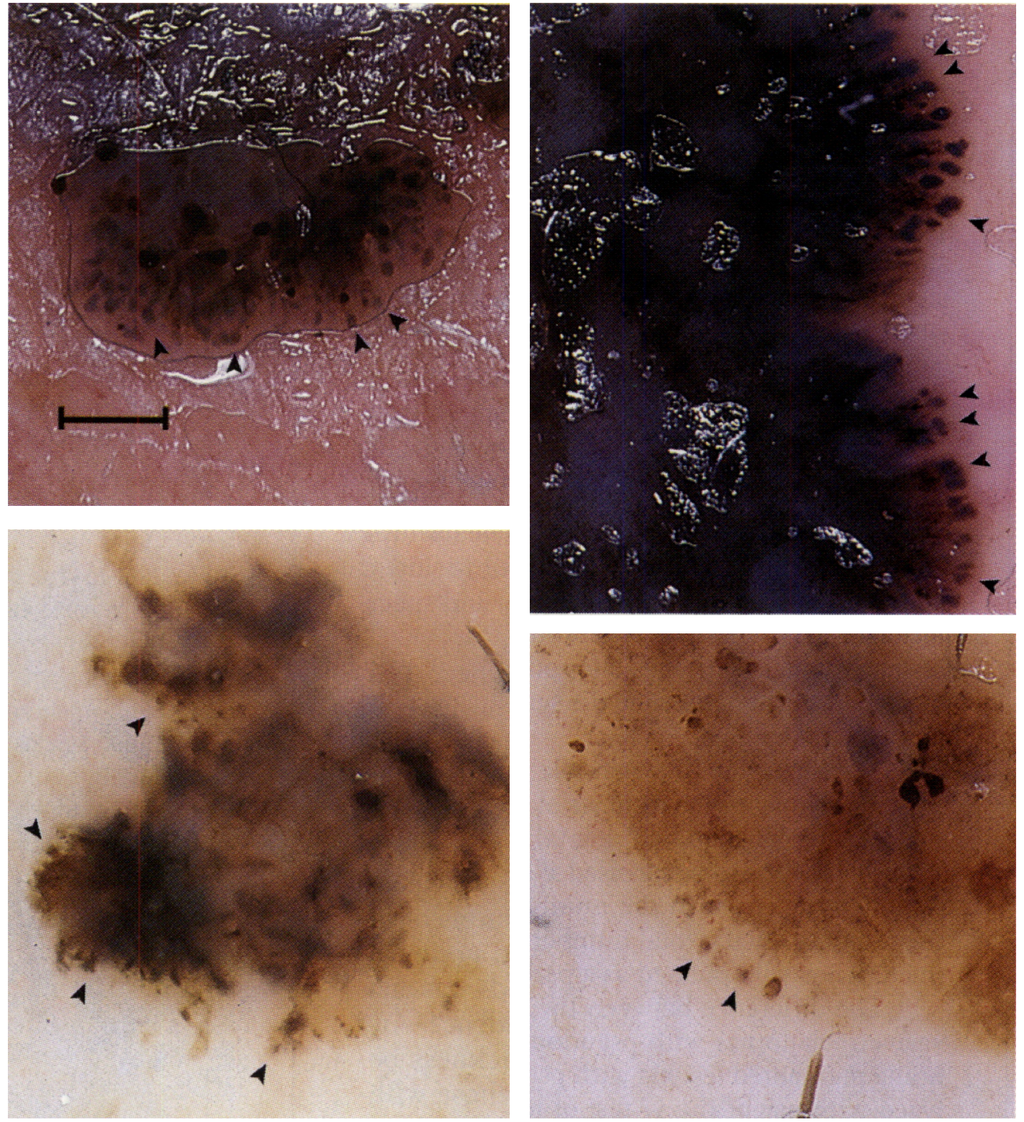
"Pseudopods are **finger-like projections of dark pigment** (brown to black) at the periphery of the lesion. They have small knobs at their tips, and are connected to either a central pigment network or central pigmented blotch." [4]



[4] Kittler et al.

Pseudopods: A Morphologic Feature in Dermoscopy

"We studied 239 pigmented lesions, 80 melanomas … the **pseudopod** retained a **97% specificity** and 23% sensitivity for **invasive melanoma**." [5]

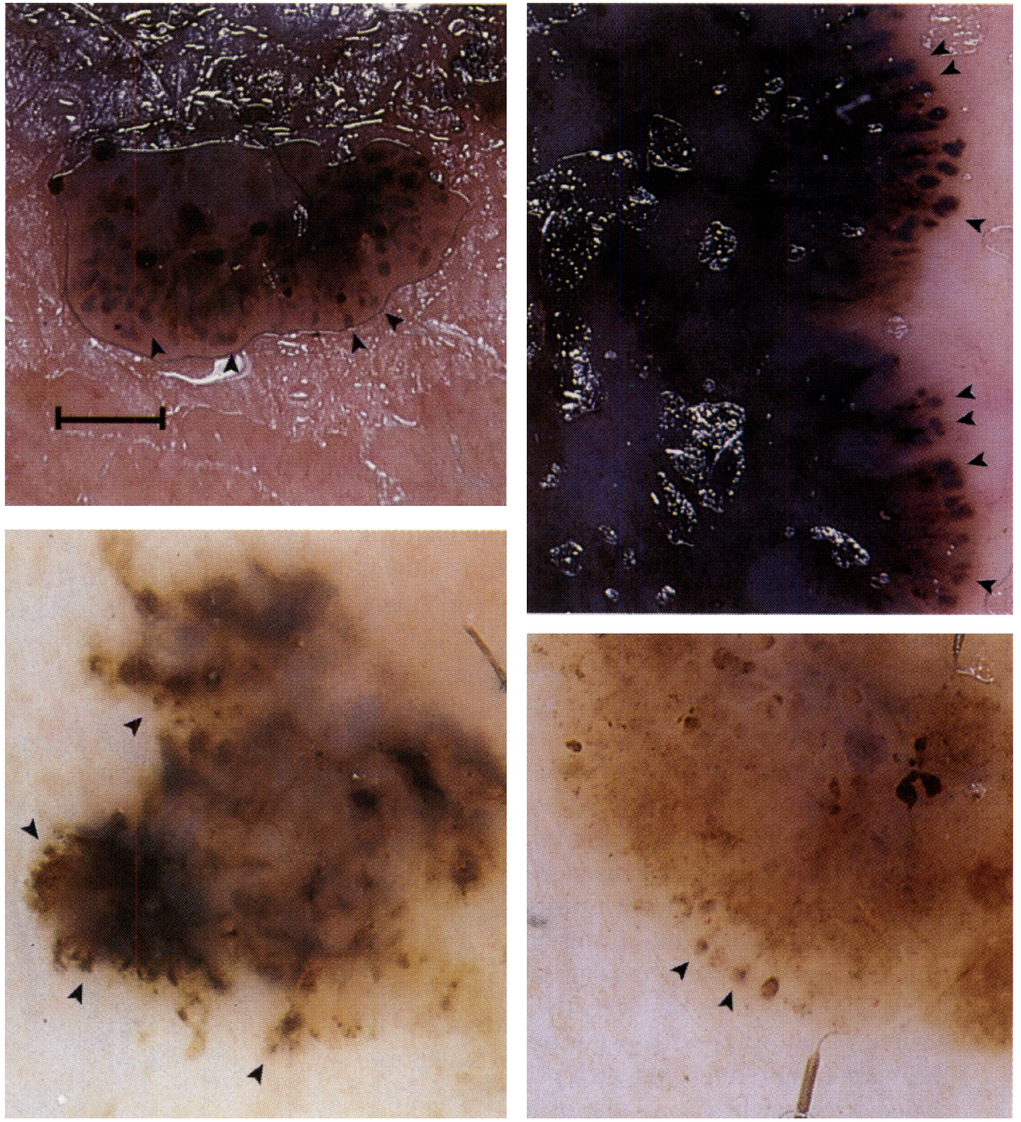


[5] Menzies et al., 1995

Pseudopods: A Morphologic Feature in Dermoscopy

"We studied 239 pigmented lesions, 80 melanomas … the **pseudopod** retained a **97% specificity** and 23% sensitivity for **invasive melanoma**." [5]

"40 studies including 22 796 skin lesions and 5736 melanomas … we affirmed the **diagnostic importance of dermoscopic structures** associated with **melanoma** detection … The features with the **highest specificity** were **pseudopods** (97.3%; 95% CI, 94.3%-98.7%) …" [6]



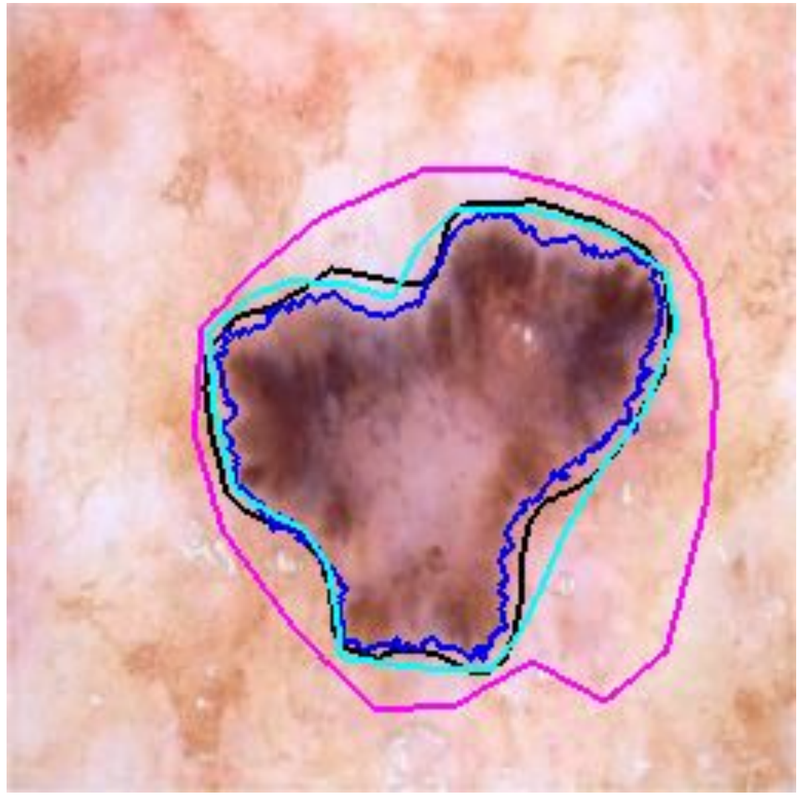
[5] Menzies et al., 1995
[6] Williams et al., 2021

Segmenting Skin Lesions with Irregular Borders

The presence of **irregular borders**, e.g., pseudopods, make it **difficult to delineate lesion borders**, and may contribute to **annotator variability**.

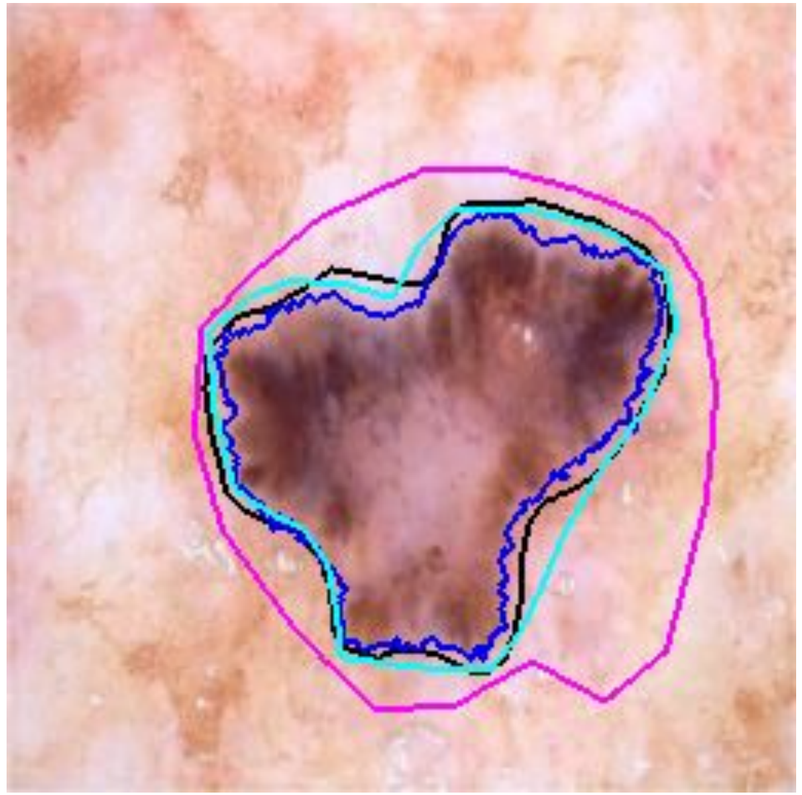
Segmenting Skin Lesions with Irregular Borders

The presence of **irregular borders**, e.g., pseudopods, make it **difficult to delineate lesion borders**, and may contribute to **annotator variability**.



Segmenting Skin Lesions with Irregular Borders

The presence of **irregular borders**, e.g., pseudopods, make it **difficult to delineate lesion borders**, and may contribute to **annotator variability**.



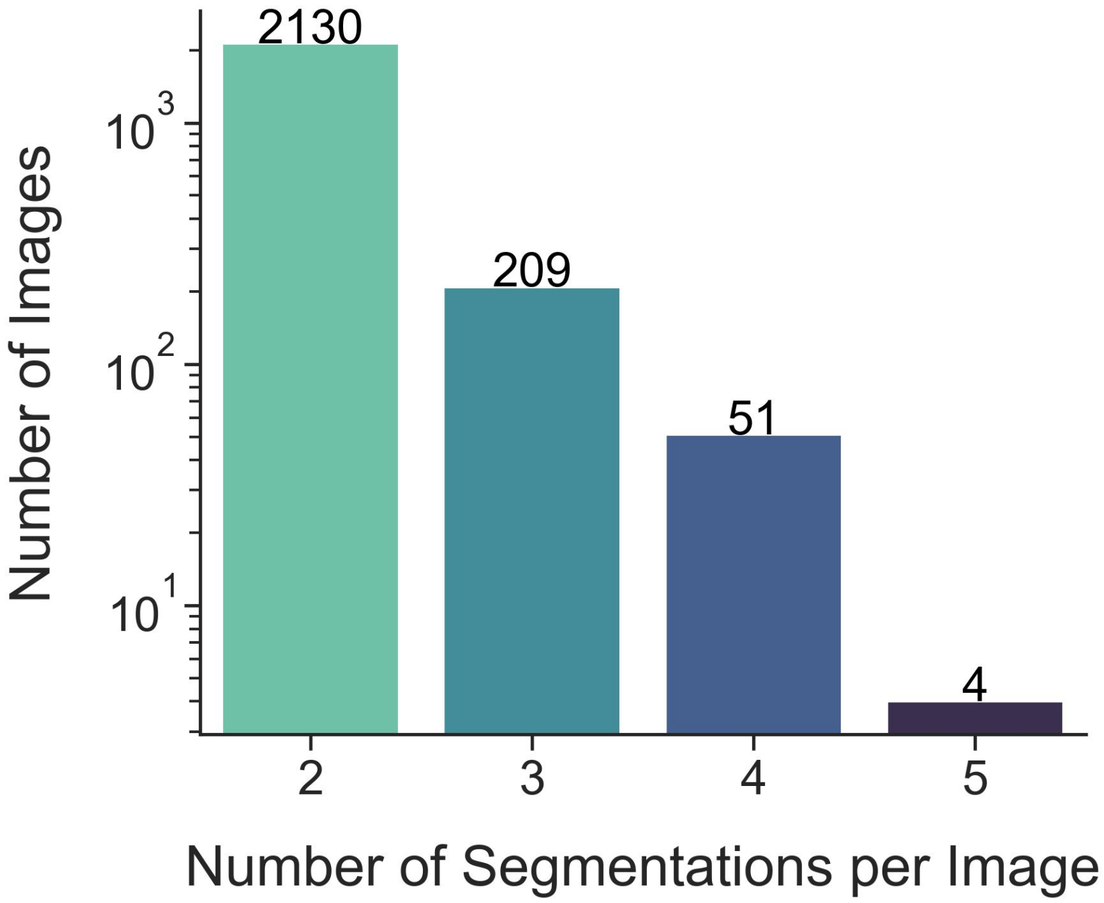
Hypothesis: Annotator (dis)agreement is related to malignancy.

No prior research investigating an association between the quantitative level of **i**nter-**a**nnotator **a**greement (**IAA**) in skin lesion segmentation and malignancy.

IMA++: A New Skin Lesion Segmentation Dataset

Curated from the ISIC Archive:

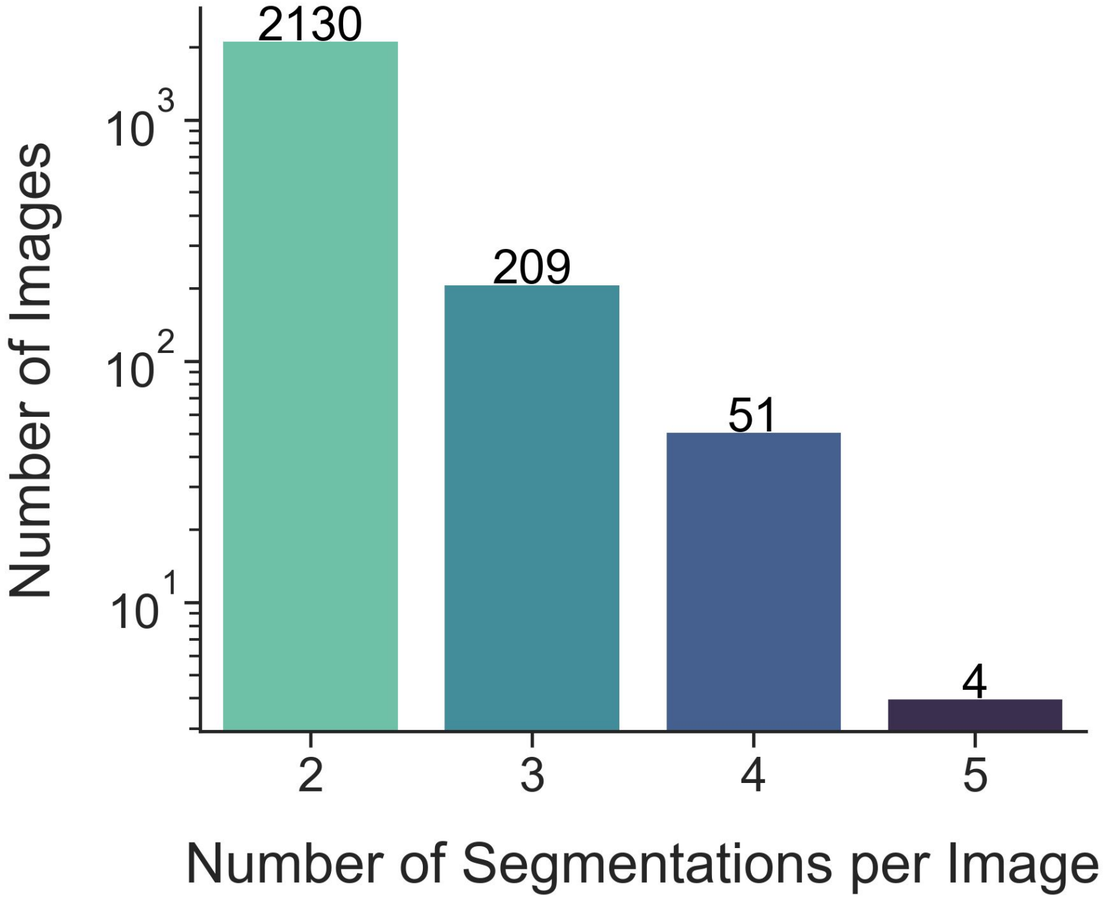
- **2,394** dermoscopic **images**
- **5,111** unique segmentation **masks**
- **15** unique **annotators**
- 3 annotation **tools**:
  - **T1:** manual polygon tracing
  - **T2:** semi-automated flood-fill
  - **T3:** fully automated seg. reviewed by expert
- 2 **skill levels**: S1, S2



IMA++: A New Skin Lesion Segmentation Dataset

Curated from the ISIC Archive:

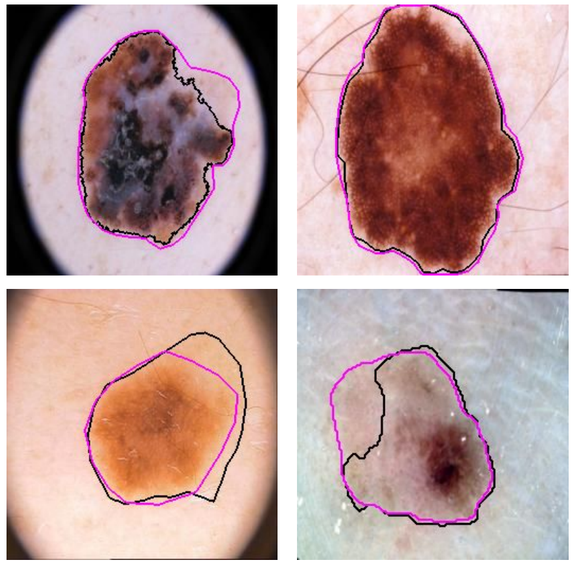
- **2,394** dermoscopic **images**
- **5,111** unique segmentation **masks**
- **15** unique **annotators**
- 3 annotation **tools**:
  - **T1:** manual polygon tracing
  - **T2:** semi-automated flood-fill
  - **T3:** fully automated seg. reviewed by expert
- 2 **skill levels**: S1, S2



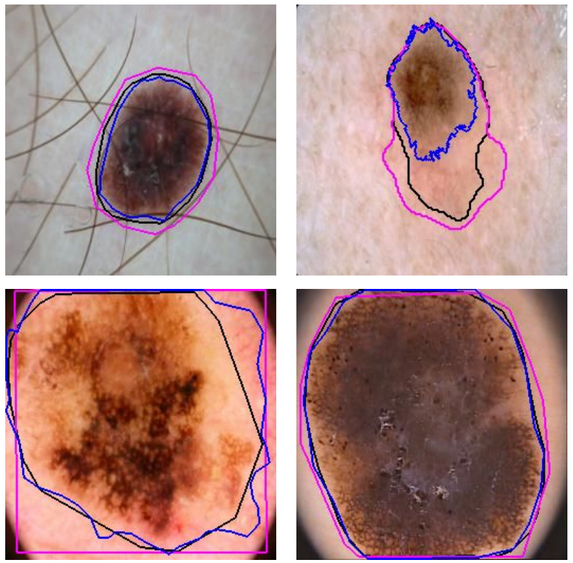
The largest public multi-annotator skin lesion segmentation dataset.

IMA++: Representative Samples

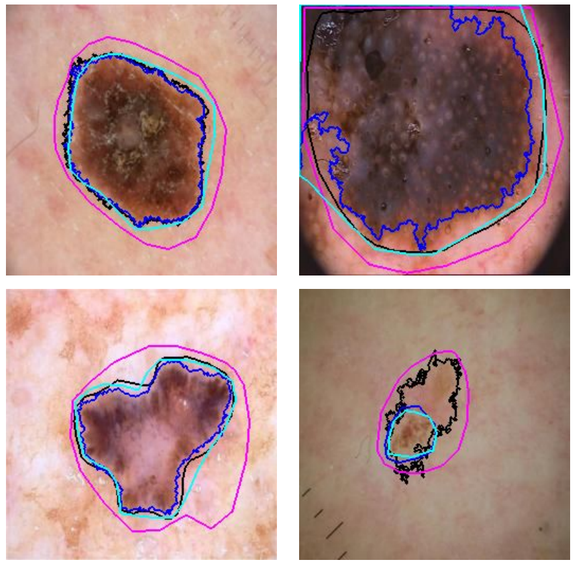
2 masks



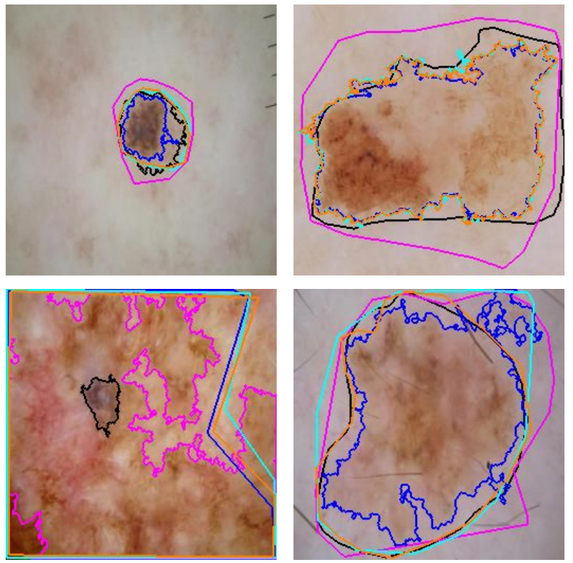
3 masks



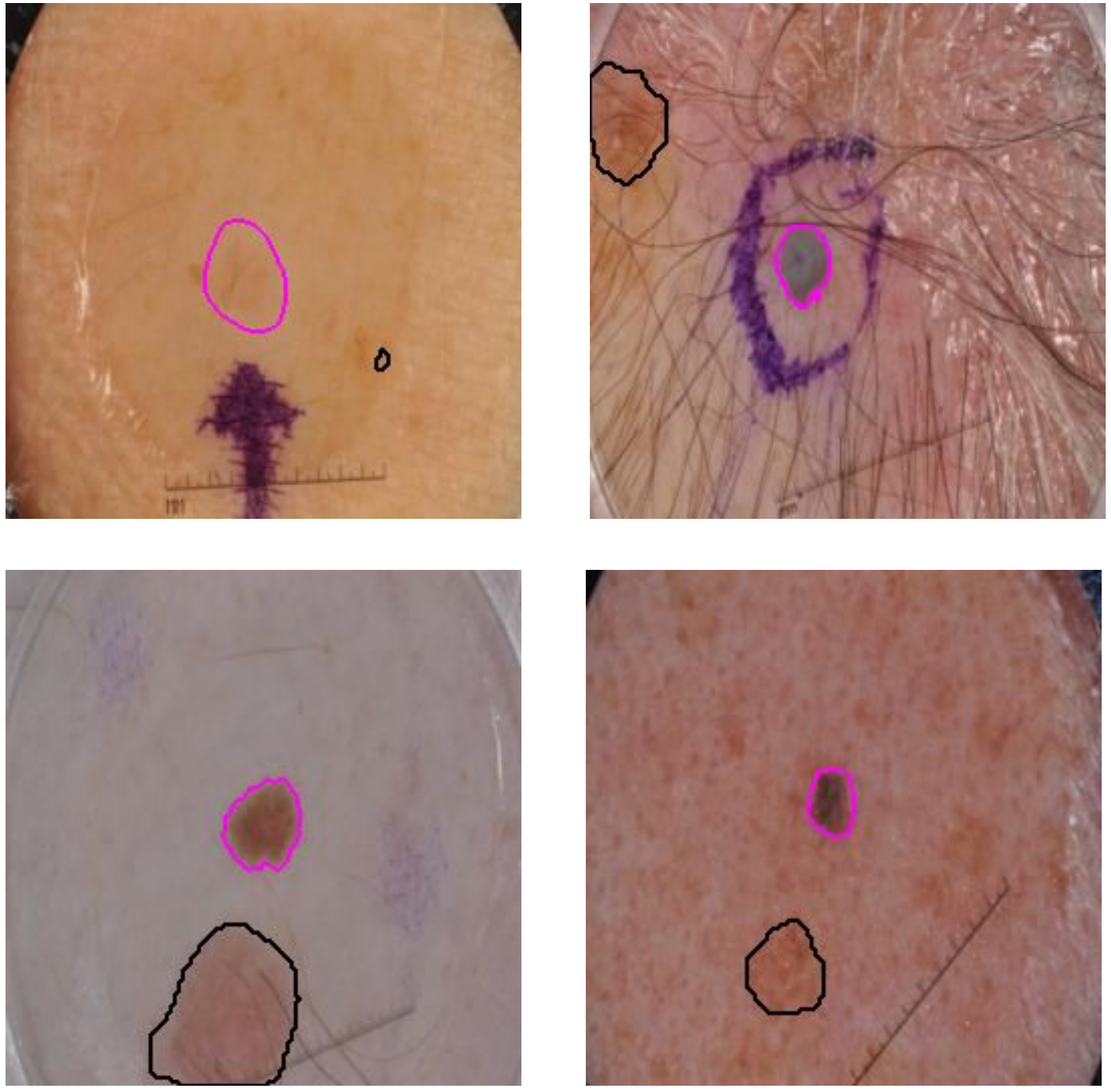
4 masks



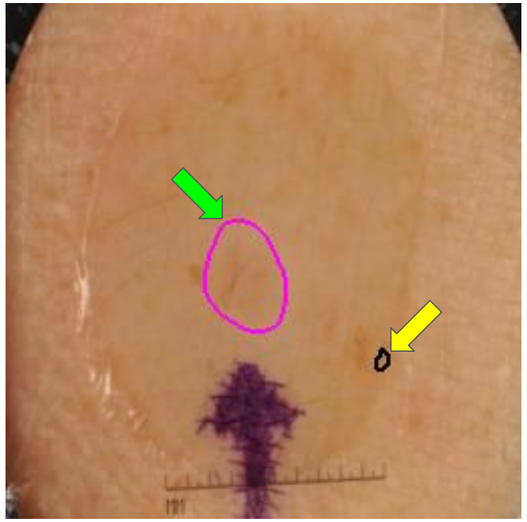
5 masks

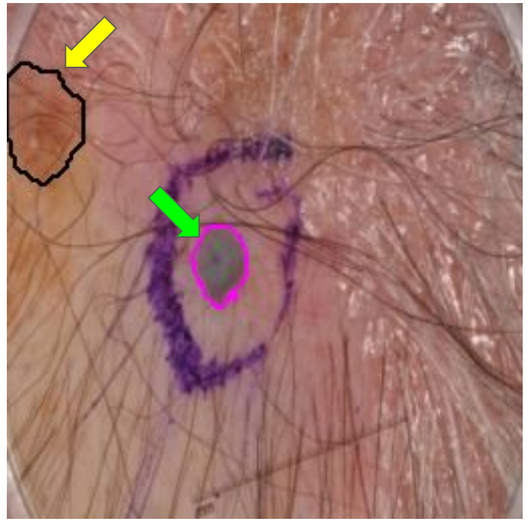


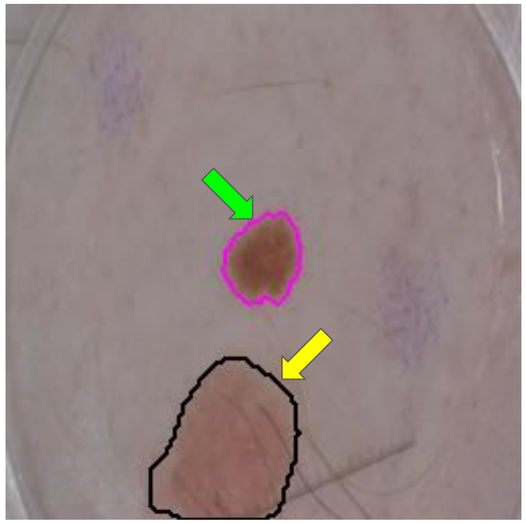
IMA++: “Conflicting” Masks

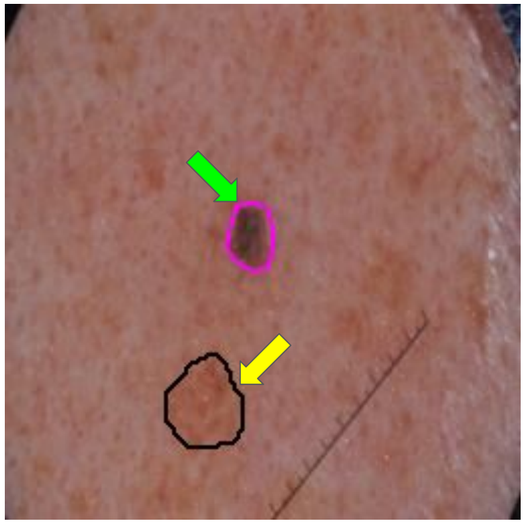


IMA++: “Conflicting” Masks

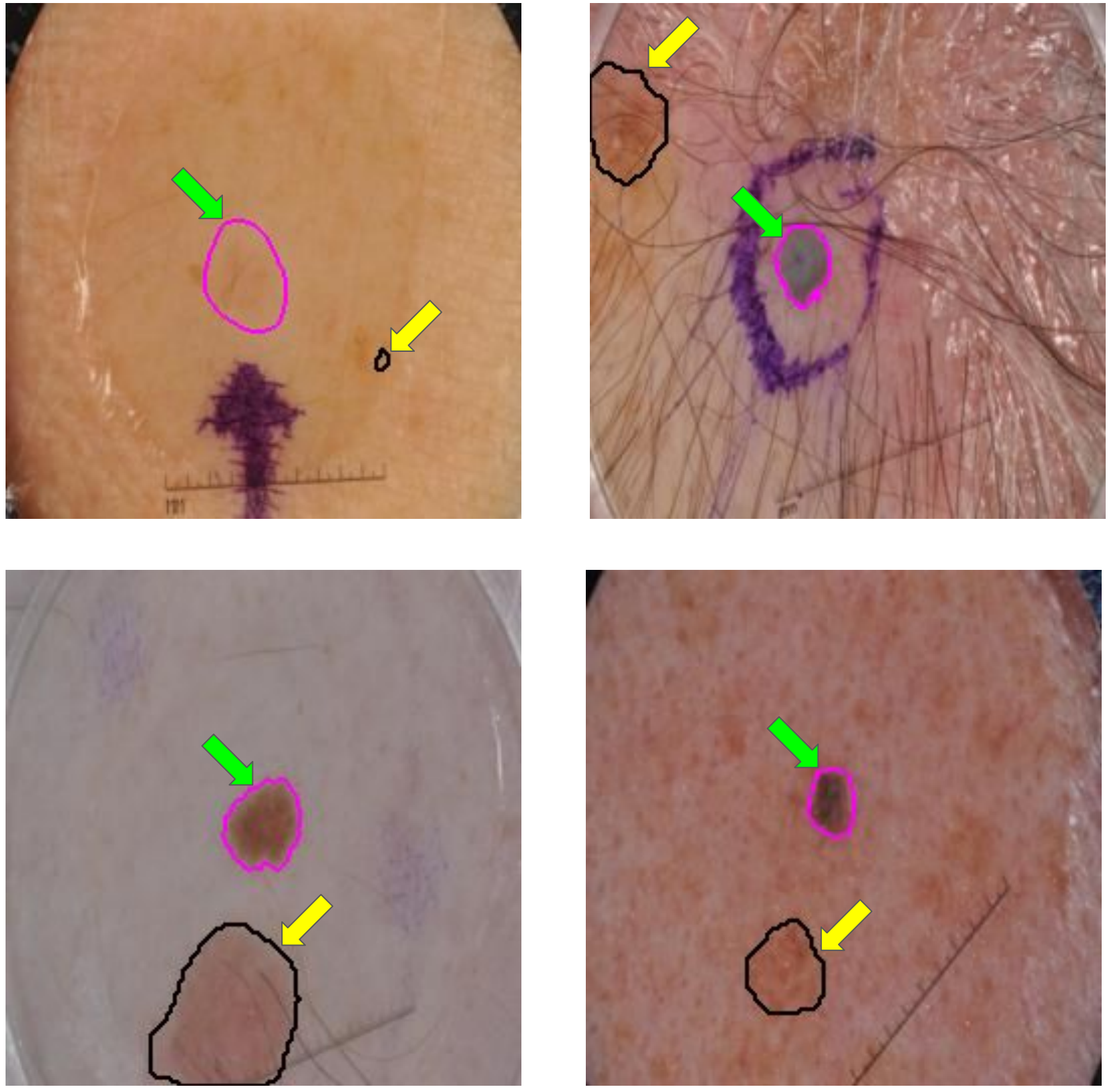








IMA++: “Conflicting” Masks



23 images have **entirely “conflicting” masks**

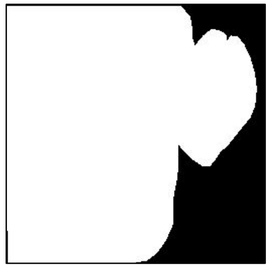
Quantifying Inter-Annotator Agreement (IAA)

For an **image $\mathbf{x_i}$** with segmentation masks **$\{\mathbf{S_{ik}}\}$**,

compute IAA score $\mathbf{Z_i} = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

ISIC_0023316



Quantifying Inter-Annotator Agreement (IAA)

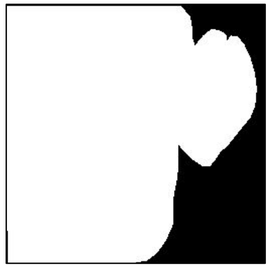
For an **image $\mathbf{x_i}$** with segmentation masks **$\{\mathbf{S_{ik}}\}$**,

compute IAA score $\mathbf{Z_i} = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

- overlap-based (Dice similarity coefficient)
- boundary-based (Hausdorff distance)

ISIC_0023316



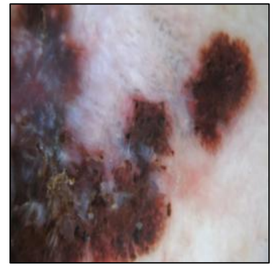
Quantifying Inter-Annotator Agreement (IAA)

For an **image x_i** with segmentation masks **$\{S_{ik}\}$**,

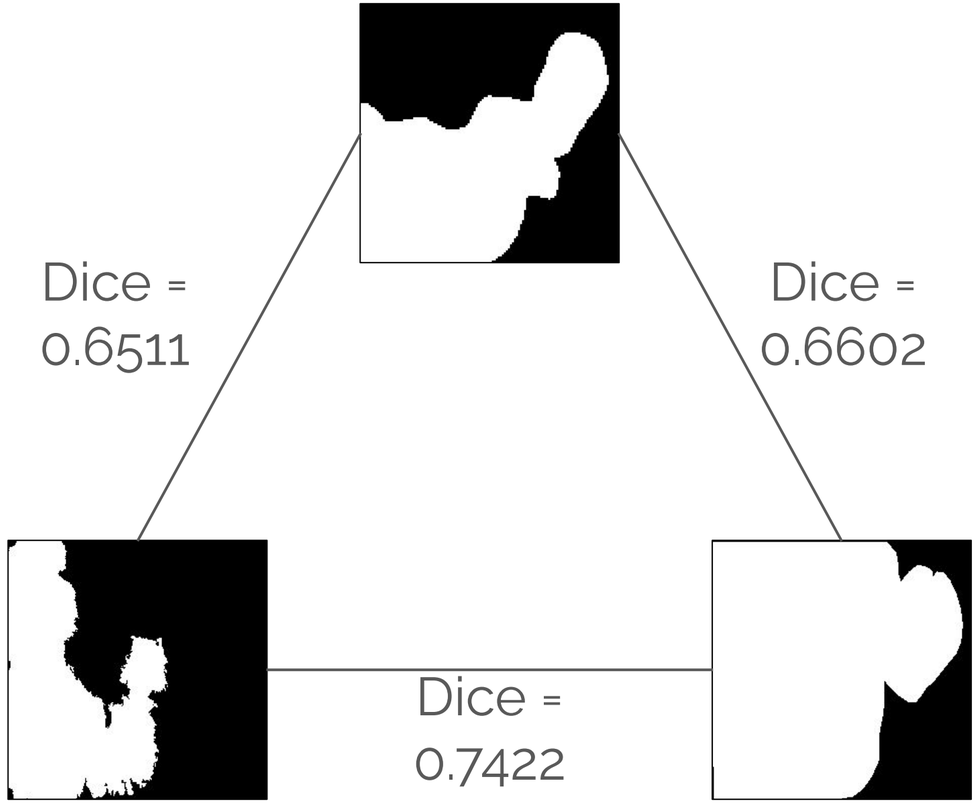
compute IAA score $Z_i = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

- overlap-based (Dice similarity coefficient)
- boundary-based (Hausdorff distance)



ISIC_0023316



Quantifying Inter-Annotator Agreement (IAA)

For an **image x_i** with segmentation masks **$\{S_{ik}\}$**,

compute IAA score $\mathbf{Z_i} = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

- overlap-based (Dice similarity coefficient)
- boundary-based (Hausdorff distance)

ISIC_0023316

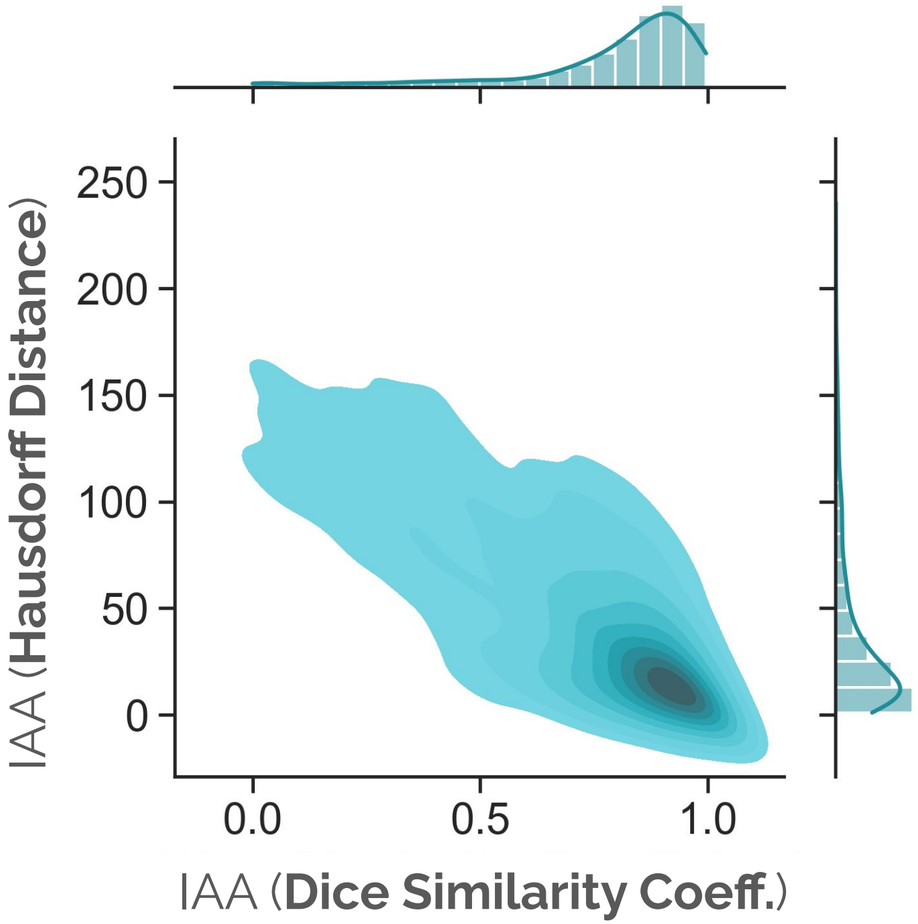
Quantifying Inter-Annotator Agreement (IAA)

For an **image $\mathbf{x_i}$** with segmentation masks **$\{S_{ik}\}$**,

compute IAA score $\mathbf{Z_i} = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

- overlap-based (Dice similarity coefficient)
- boundary-based (Hausdorff distance)



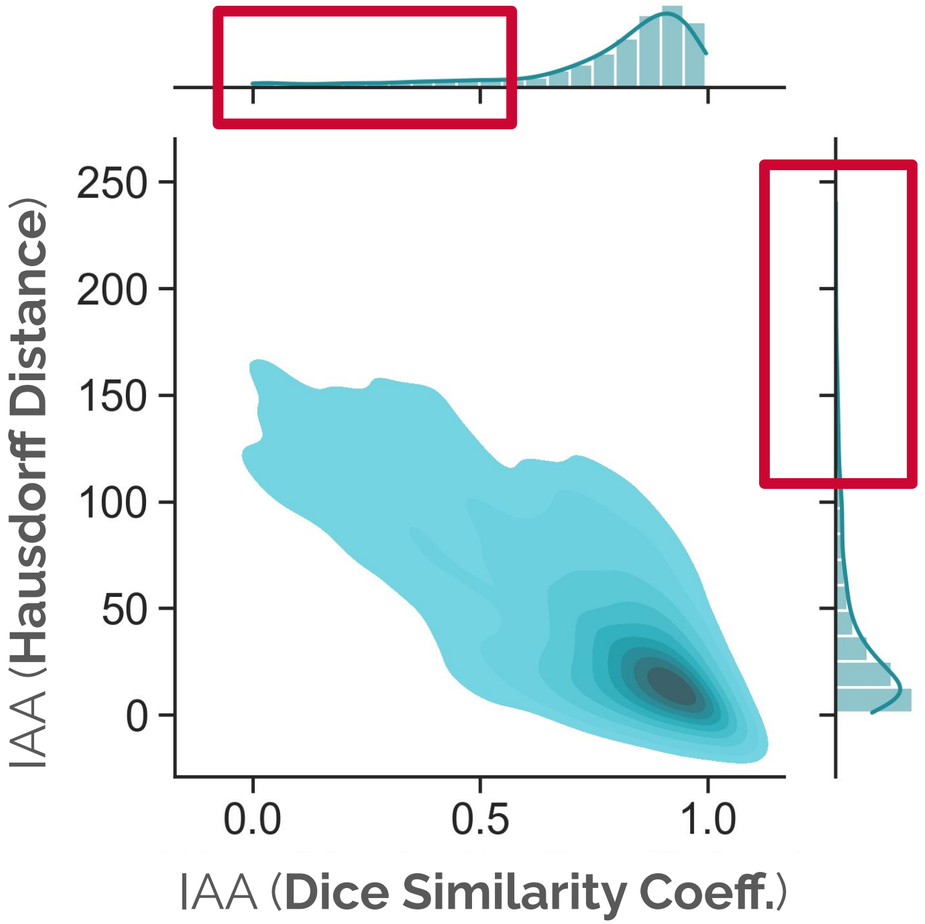
Quantifying Inter-Annotator Agreement (IAA)

For an **image x_i** with segmentation masks **$\{S_{ik}\}$**,

compute IAA score $\mathbf{Z_i} = g(\{S_{ik}\})$,

where $g()$ is a similarity measure:

- overlap-based (Dice similarity coefficient)
- boundary-based (Hausdorff distance)



Skewed distributions

Peaks at high IAA
Long tails extending to 0 IAA

Inter- and Intra-Factor Agreement in IMA++

Factors: annotator, segmentation tool, skill, lesion malignancy.

Inter- and Intra-Factor Agreement in IMA++

Factors: annotator, segmentation tool, skill, lesion malignancy.

Analysis:

- **Mann-Whitney *U* Test:** assess if the factor-based differences are stat. sig.

- **Cohen's *d*:** quantifies the effect size to show the magnitude of the difference.

High Intra-Annotator Agreement

Factors: annotator, segmentation tool, skill, lesion malignancy.

Annotators agree more with themselves than they do with others.

	Annotator	
	Same	Different
IAA	0.900 ± 0.131	0.772 ± 0.221
p-value	1.85E-35	
Cohen's *d*	2.714	

Tool(s) Used and Annotator Skill Level Matter

Factors: annotator, segmentation tool, skill, lesion malignancy.

Annotators agree more when they use the same tool or have similar skill levels.

	Tool		Skill	
	Same	Different	Same	Different
IAA	0.862 ± 0.157	0.747 ± 0.231	0.806 ± 0.167	0.710 ± 0.258
p-value	2.45E-69		1.17E-05	
Cohen's *d*	2.447		1.816	

Lesion Malignancy Significantly Affects IAA

Factors: annotator, segmentation tool, skill, lesion malignancy.

Malignant skin lesions tend to exhibit **lower IAA** (Dice).

	Malignancy	
	Benign	Malignant
IAA	0.791 ± 0.225	0.753 ± 0.227
p-value	4.77E-06	
Cohen's *d*	0.798	

Conclusion: Lesion boundary ambiguity captured by IAA aligns with malignancy.

Testing for Systematic Difference in IAA Distributions

First order stochastic dominance **(FOSD)** test to examine if a systematic difference exists between IAA scores for benign and malignant lesions.

Testing for Systematic Difference in IAA Distributions

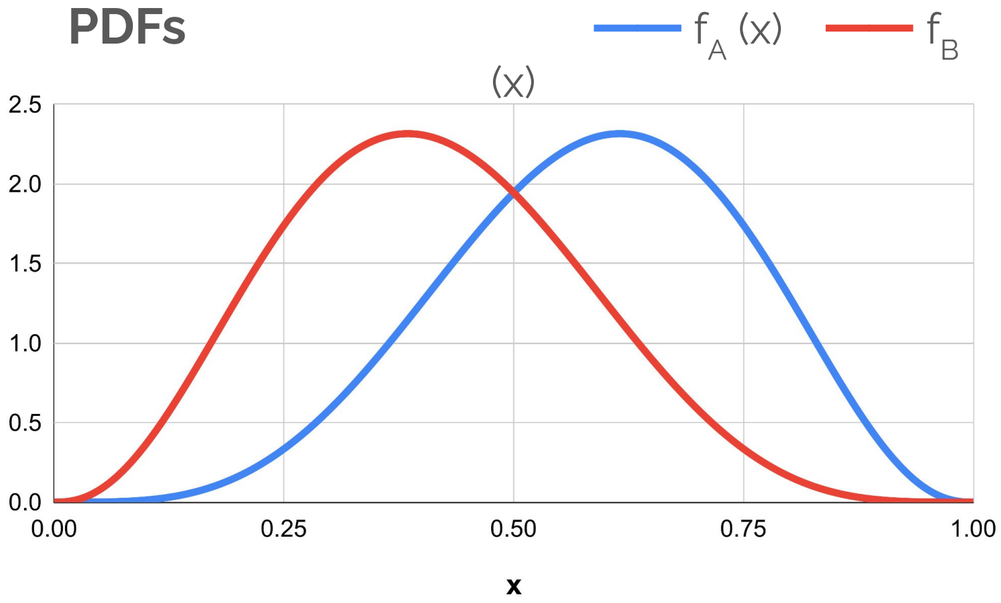
First order stochastic dominance **(FOSD)** test to examine if a systematic difference exists between IAA scores for benign and malignant lesions.

FOSD: dist. $f_A(x)$ first-order stochastically dominates dist. $f_B(x)$ if $\forall x$, with strict inequality for some x:

$$\mathbf{F_A(x) \leq F_B(x)}.$$

Testing for Systematic Difference in IAA Distributions

First order stochastic dominance **(FOSD)** test to examine if a systematic difference exists between IAA scores for benign and malignant lesions.



FOSD: dist. $f_A(x)$ first-order stochastically dominates dist. $f_B(x)$ if $\forall x$, with strict inequality for some x:

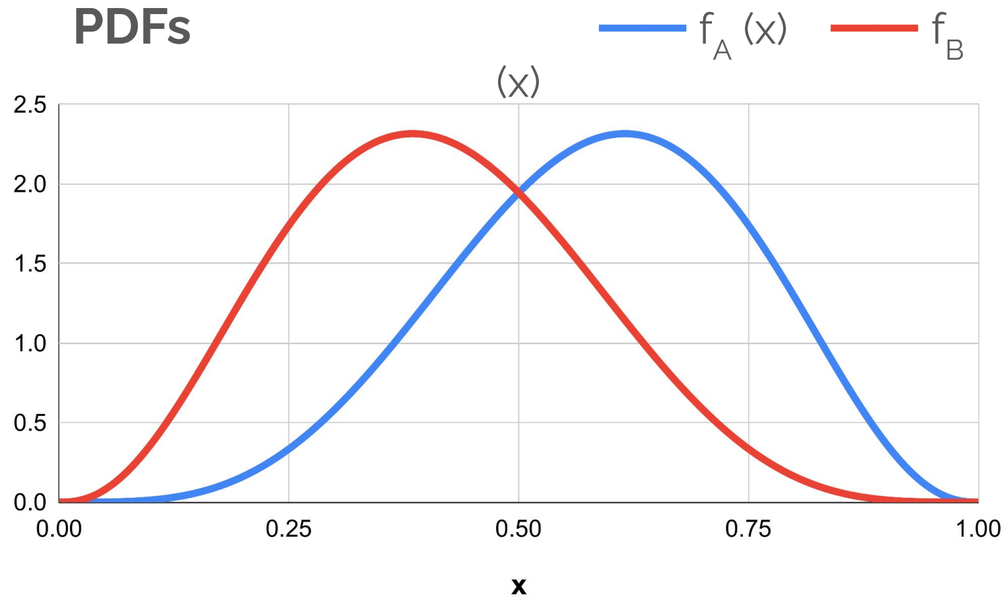
$$\mathbf{F_A(x) \leq F_B(x)}.$$

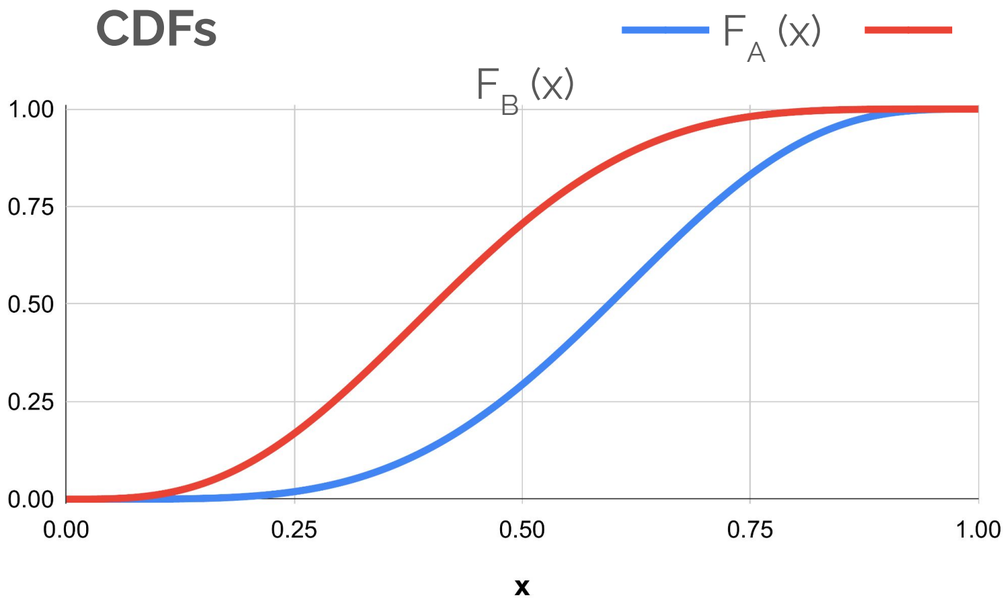
Testing for Systematic Difference in IAA Distributions

First order stochastic dominance **(FOSD)** test to examine if a systematic difference exists between IAA scores for benign and malignant lesions.

FOSD: dist. $f_A(x)$ first-order stochastically dominates dist. $f_B(x)$ if $\forall x$, with strict inequality for some x:

$$\mathbf{F_A(x) \leq F_B(x)}.$$





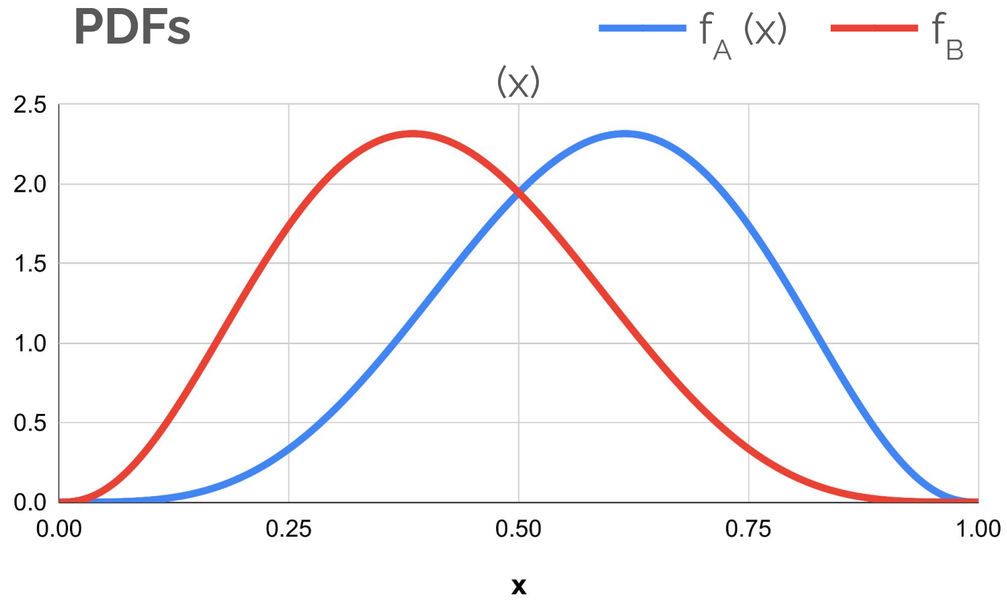
Testing for Systematic Difference in IAA Distributions

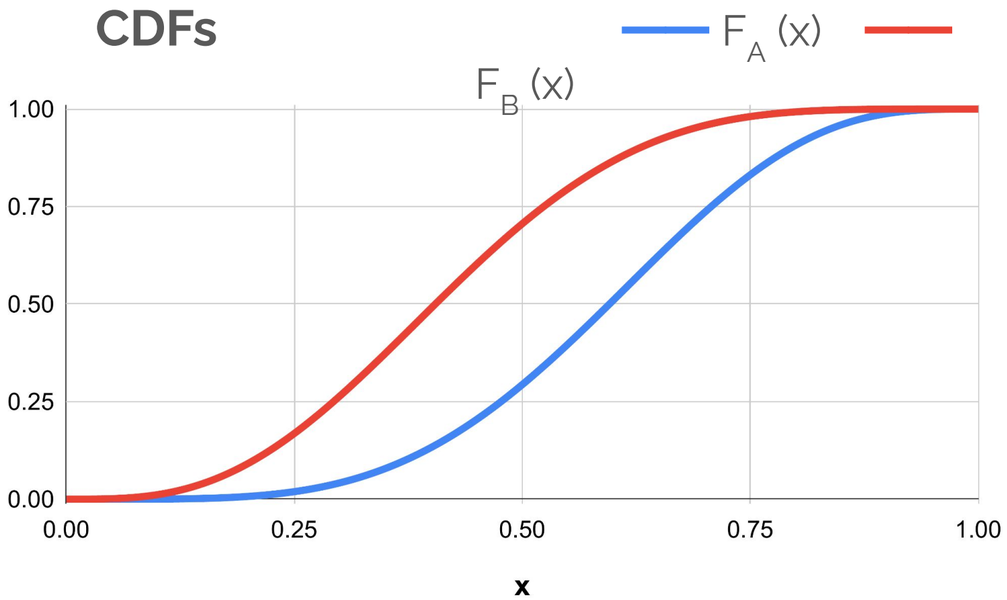
First order stochastic dominance **(FOSD)** test to examine if a systematic difference exists between IAA scores for benign and malignant lesions.

FOSD: dist. $f_A(x)$ first-order stochastically dominates dist. $f_B(x)$ if $\forall x$, with strict inequality for some x:

$$\mathbf{F_A(x) \leq F_B(x)}.$$

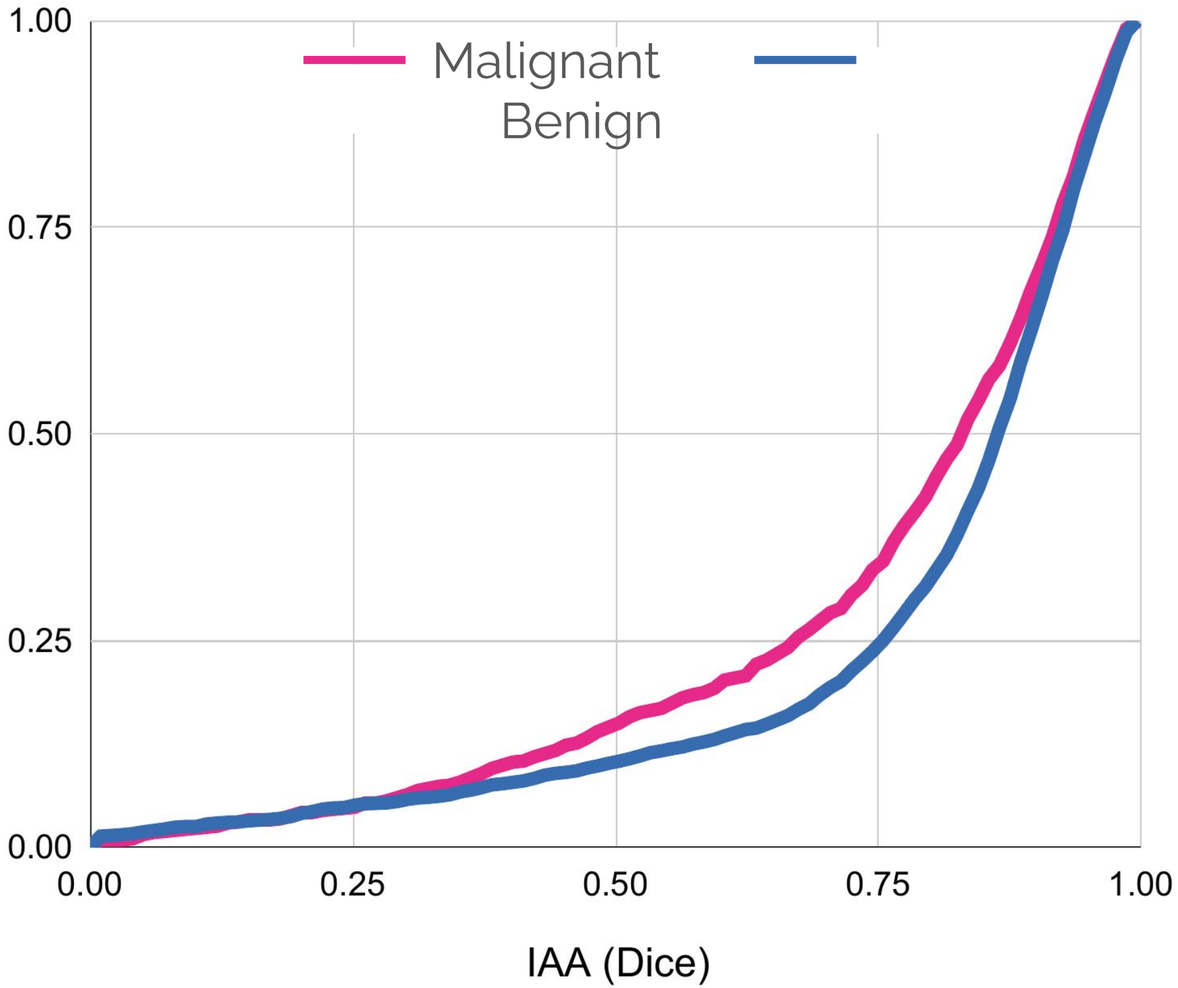
This is denoted by $F_A \geq_1 F_B$.





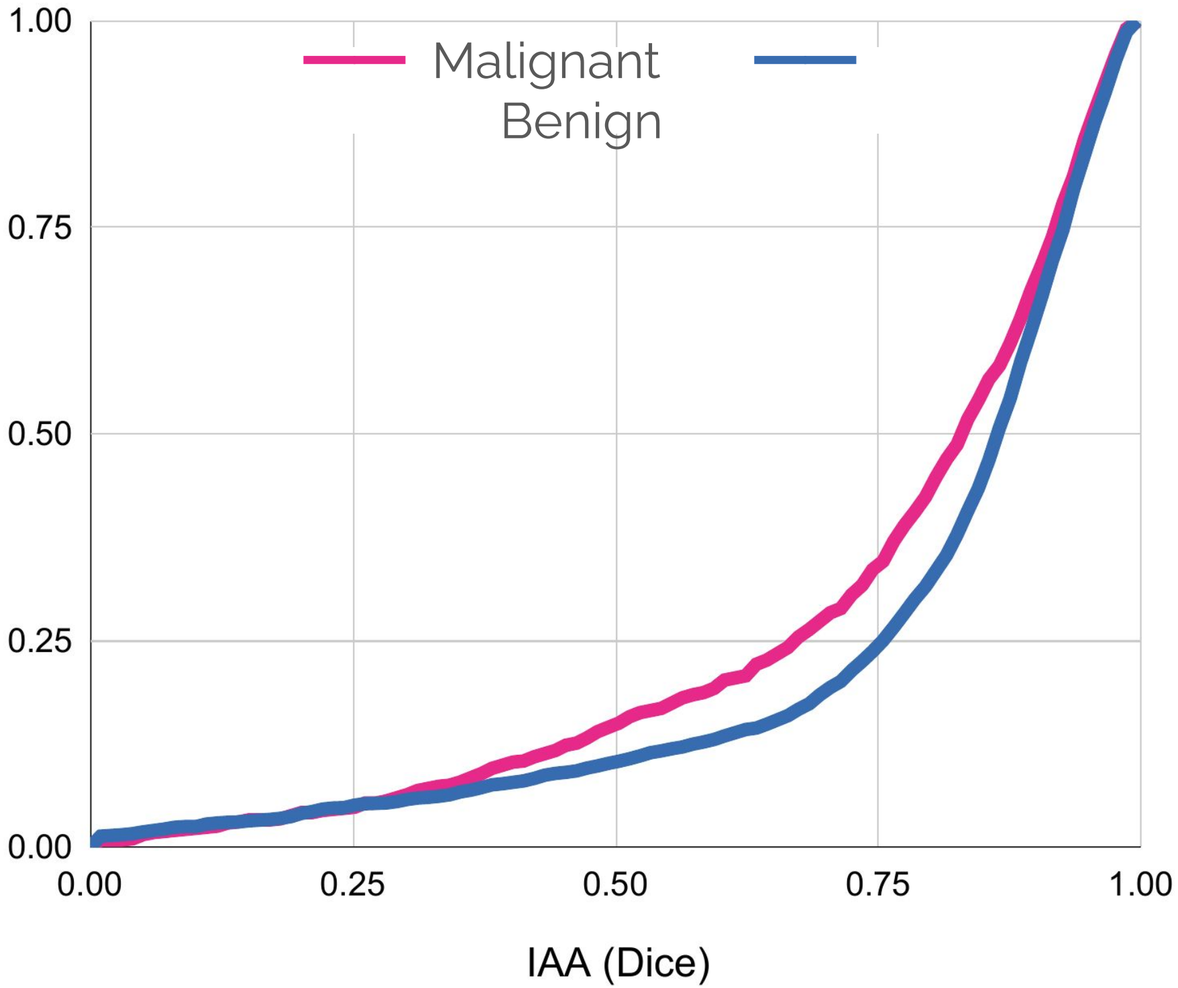
IAA Distribution Shifts due to Malignancy

CDFs of **Malignant** vs **Benign**



IAA Distribution Shifts due to Malignancy

CDFs of **Malignant** vs **Benign**



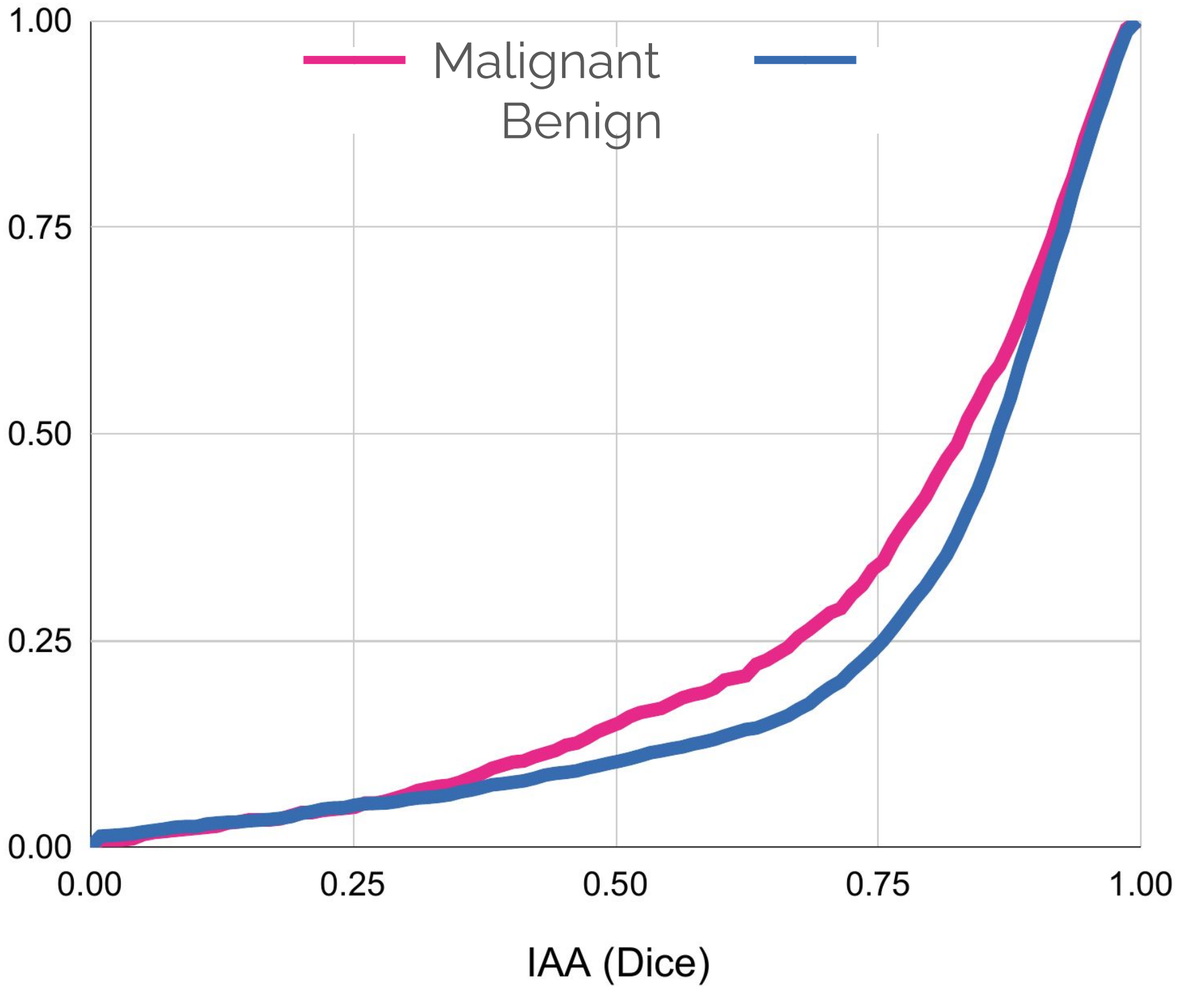
$F_{ben} \leq F_{mal}$?

Two one-sided FOSD tests:

- $H_{mal \geq_1 ben}$ →rejected
- $H_{ben \geq_1 mal}$ →failed to reject

IAA Distribution Shifts due to Malignancy

CDFs of **Malignant** vs **Benign**



$F_{ben} \leq F_{mal}$?

Two one-sided FOSD tests:

- $H_{mal \geq_1 ben}$ →rejected
- $H_{ben \geq_1 mal}$ →failed to reject

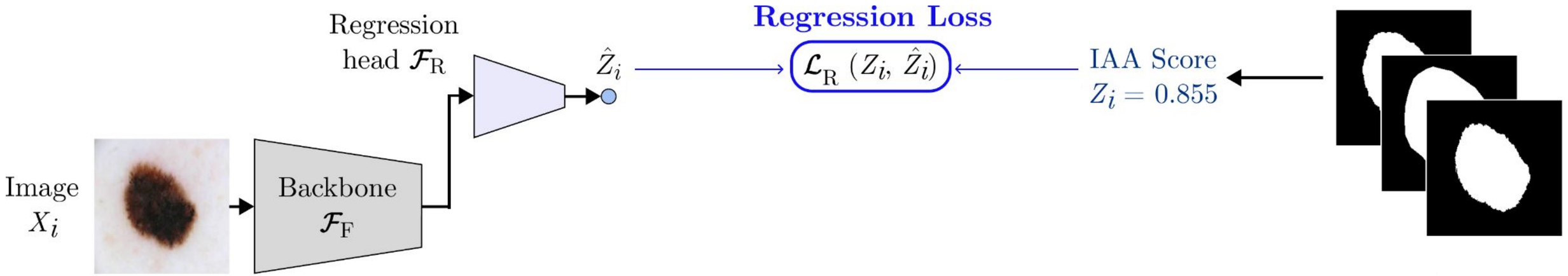
Benign lesions exhibit higher segmentation consensus.

Can We Predict IAA from Images Alone?

Given a skin lesion image X_i, can we directly predict Z_i without requiring access to the underlying segmentations?

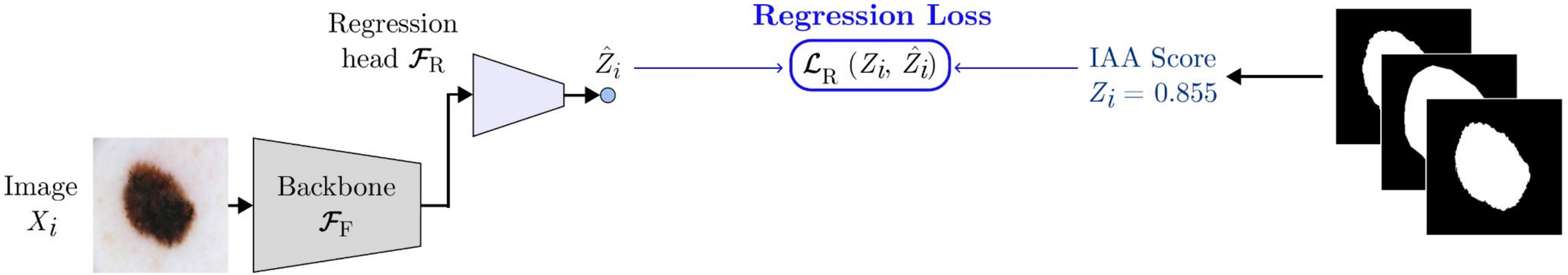
Can We Predict IAA from Images Alone?

Given a skin lesion image X_i, can we directly predict Z_i without requiring access to the underlying segmentations?



Can We Predict IAA from Images Alone?

Given a skin lesion image X_i, can we directly predict Z_i without requiring access to the underlying segmentations?

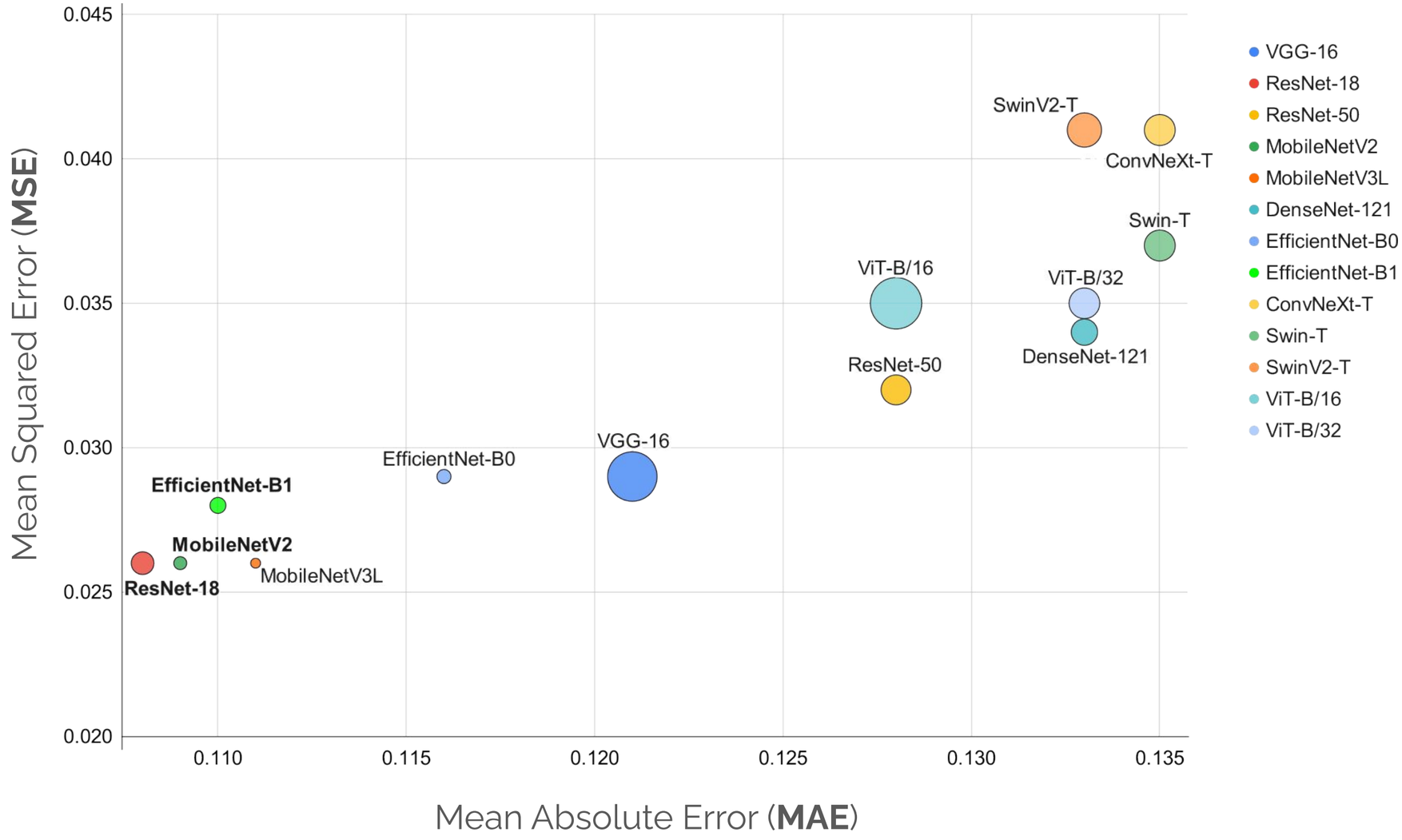


Experiments:

- 13 CNN & ViT backbones with a **regression head**.
- **SmoothL1 loss** (L1 loss for large errors; L2 loss for small errors).
- MAE and MSE reported; model with **best MAE** chosen.

IAA Can Be Predicted from Images Alone

13 models of varying compute sizes (multiply-accumulate operations; **MACs**).



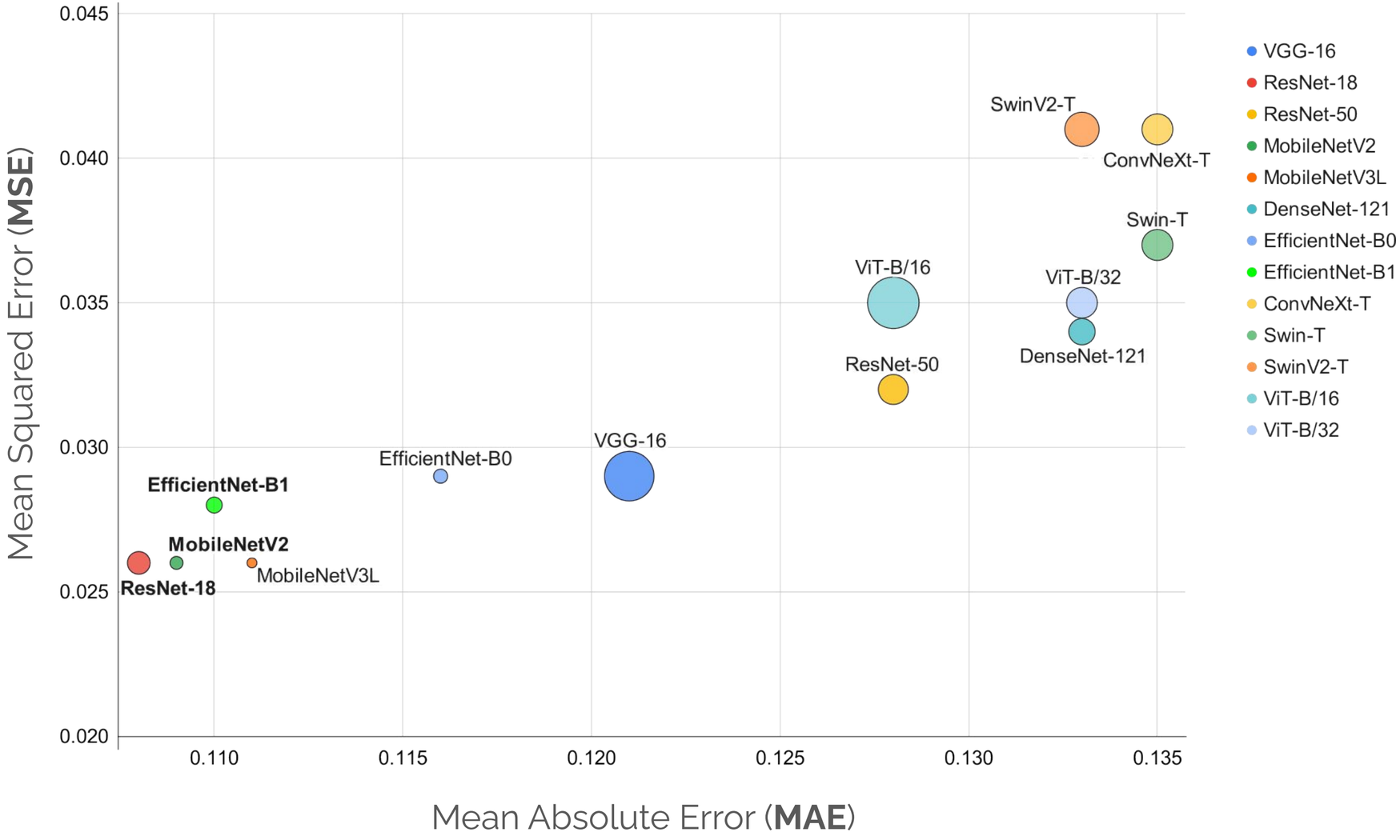
IAA Can Be Predicted from Images Alone

13 models of varying compute sizes (multiply-accumulate operations; **MACs**).

All models predict MAE in [0.108, 0.135].

Top 3 models:

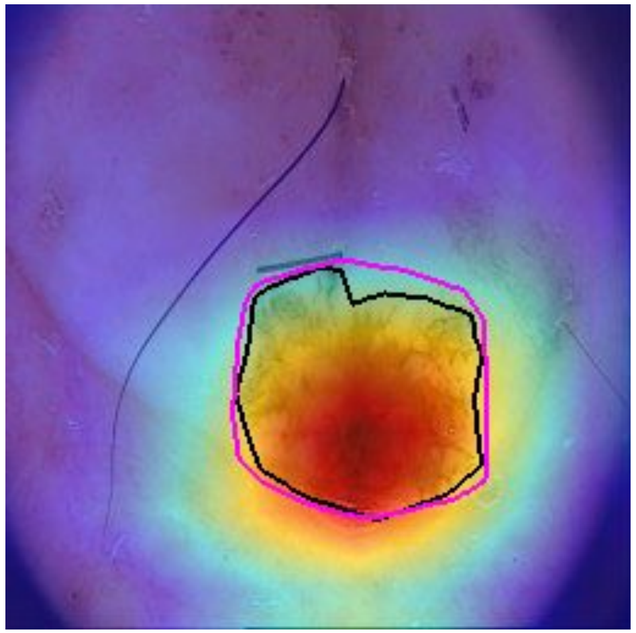
- ResNet-18 (MAE = 0.108)
- MobileNetV2 (MAE = 0.109)
- EfficientNet-B1 (MAE = 0.110)



IAA Regressor Learns to Localize Lesion Boundary

Grad-CAM++ heatmaps for the ResNet-18 regressor show that the model learns **lesion boundary ambiguity cues** that reflect annotator variability.

ISIC_0000373



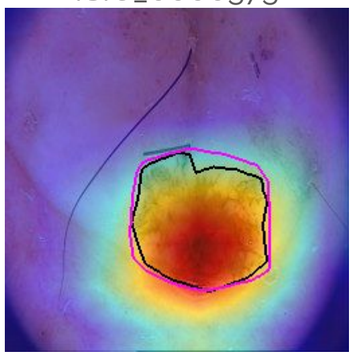
True IAA 0.9197

Pred. IAA 0.8823

IAA Regressor Learns to Localize Lesion Boundary

Grad-CAM++ heatmaps for the ResNet-18 regressor show that the model learns **lesion boundary ambiguity cues** that reflect annotator variability.

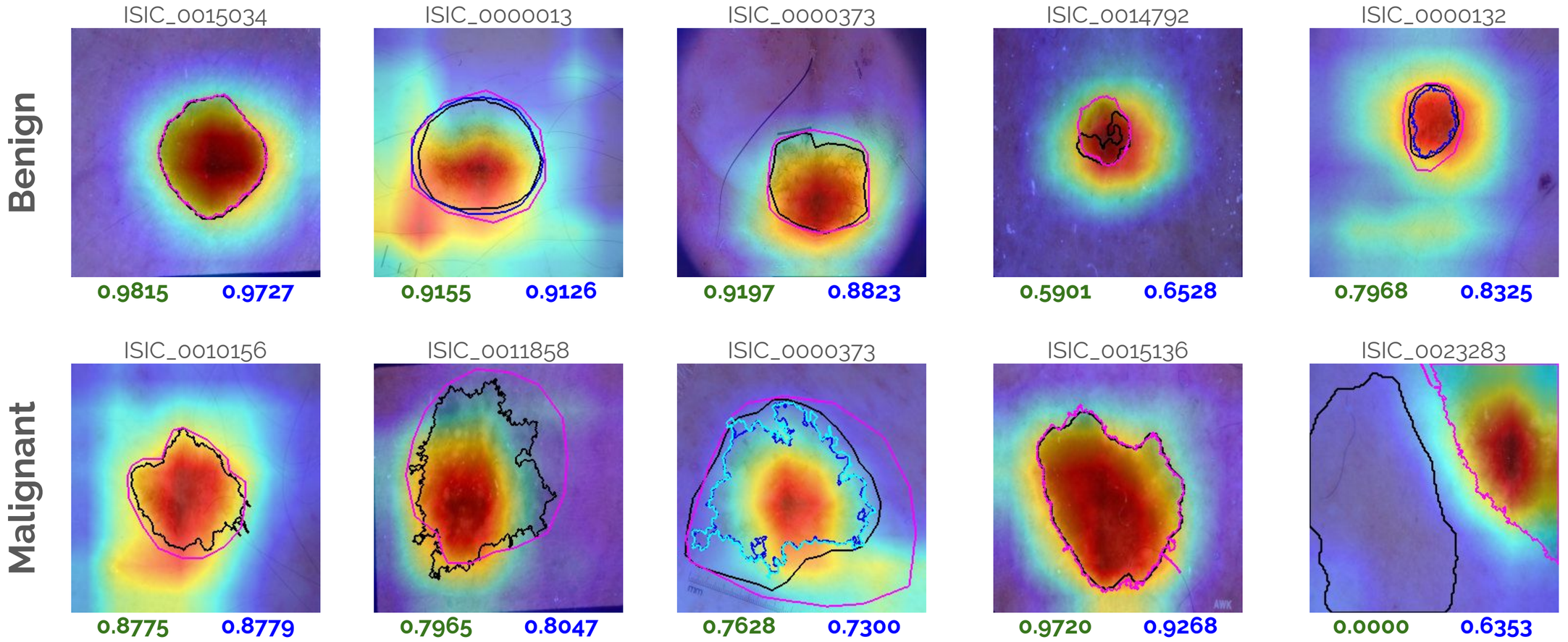
ISIC_0000373



0.9197 0.8823

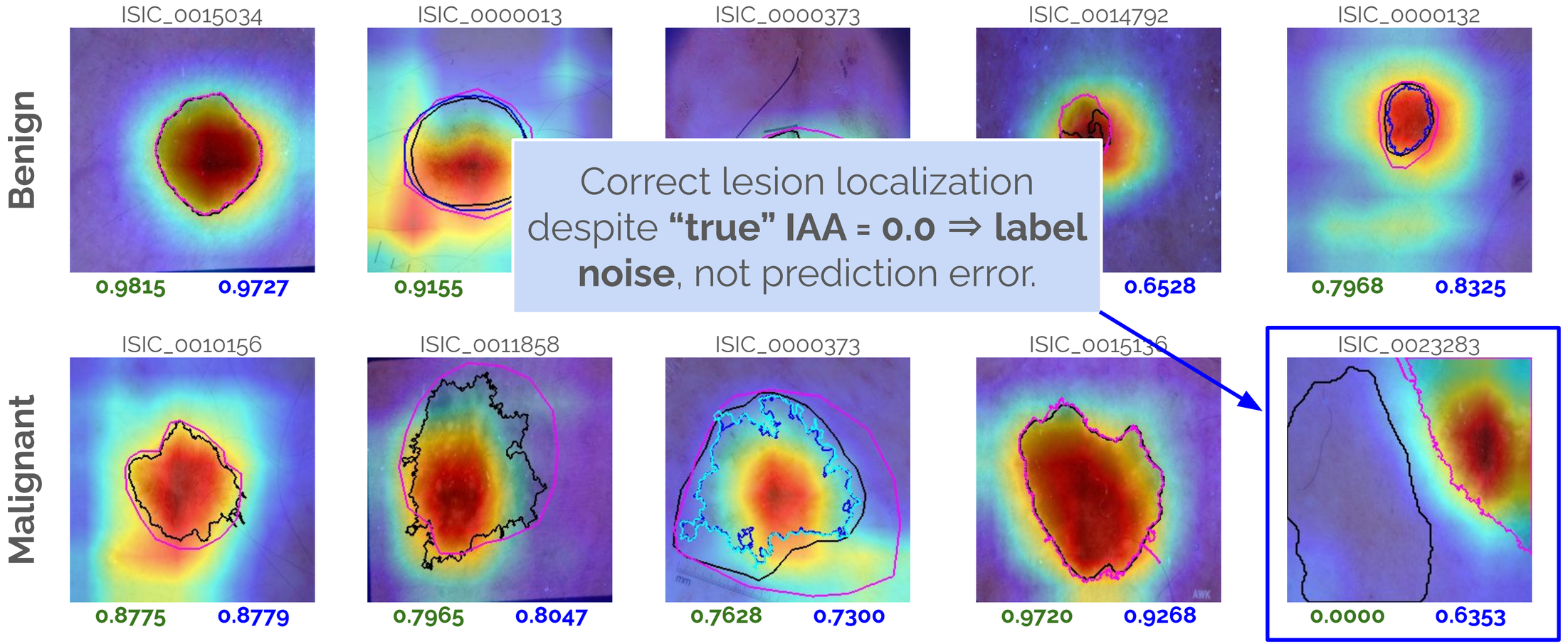
IAA Regressor Learns to Localize Lesion Boundary

Grad-CAM++ heatmaps for the ResNet-18 regressor show that the model learns **lesion boundary ambiguity cues** that reflect annotator variability.



IAA Regressor Learns to Localize Lesion Boundary

Grad-CAM++ heatmaps for the ResNet-18 regressor show that the model learns **lesion boundary ambiguity cues** that reflect annotator variability.



Can We Leverage IAA as a “Soft” Clinical Feature?

Multi-task methods (diagnosis + segmentation) improve diagnosis performance

But, lesion segmentation can be affected by inter-annotator differences.

Can We Leverage IAA as a “Soft” Clinical Feature?

Multi-task methods (diagnosis + segmentation) improve diagnosis performance

But, lesion segmentation can be affected by inter-annotator differences.

Hypothesis: Learning the variability in human interpretation inherently captures complex morphological characteristics indicative of malignancy (e.g., border irregularity, asymmetry), which are often difficult to formalize/influenced by annotator subjectivity.

Research Question: Does simultaneous prediction of IAA and diagnosis improve the latter?

Predicting IAA and Diagnosis in a Multi-Task Framework

A multi-task model:

- **Regression head** → IAA
- **Classification head** → diagnosis

- Multi-task **loss**: $\alpha L_{\text{diagnosis}} + (1 - \alpha) L_{\text{regression}}$

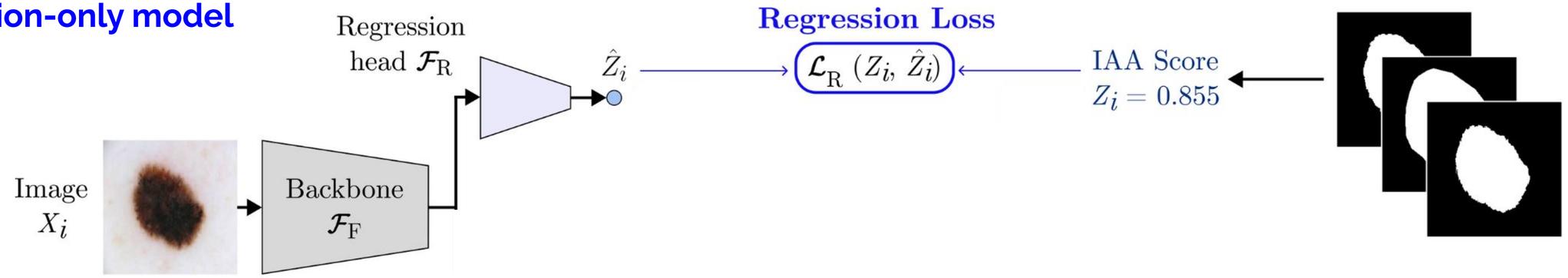
Predicting IAA and Diagnosis in a Multi-Task Framework

A multi-task model:

- **Regression head** → IAA
- **Classification head** → diagnosis

- Multi-task **loss**: $\alpha L_{\text{diagnosis}} + (1 - \alpha) L_{\text{regression}}$

From the original
regression-only model



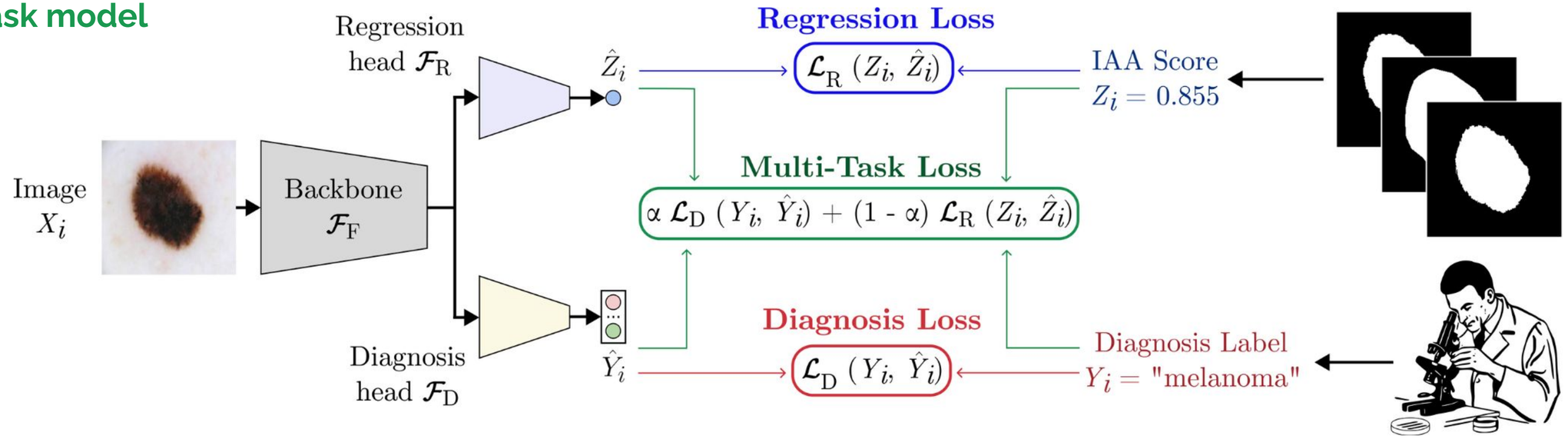
Predicting IAA and Diagnosis in a Multi-Task Framework

A multi-task model:

- **Regression head** → IAA
- **Classification head** → diagnosis

- Multi-task **loss**: $\alpha L_{\text{diagnosis}} + (1 - \alpha) L_{\text{regression}}$

To a new
multi-task model



How do Multi-Task Models Fare Against Diag. Only Models?

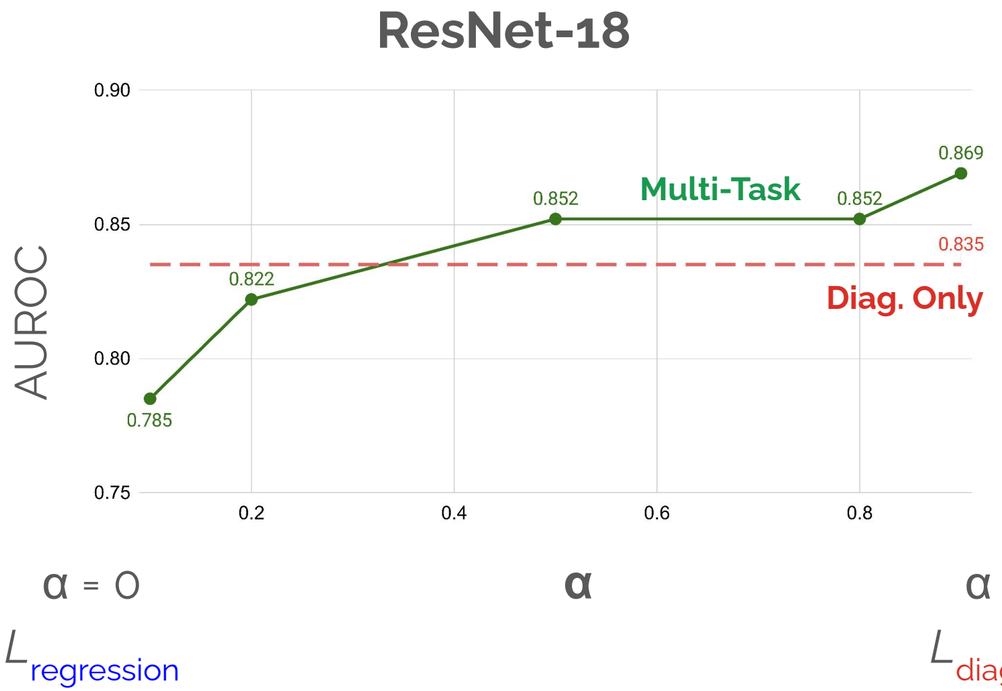
Experiment: Vary α to study the relative importance of IAA prediction.

Multi-Task Models Diagnose Better than Diag. Only Models

Experiment: Vary α to study the relative importance of IAA prediction.

Results:

- Diagnosis-dominant (α = 0.9) multi-task models perform the best.
- Multi-task models outperform diagnosis-only models.

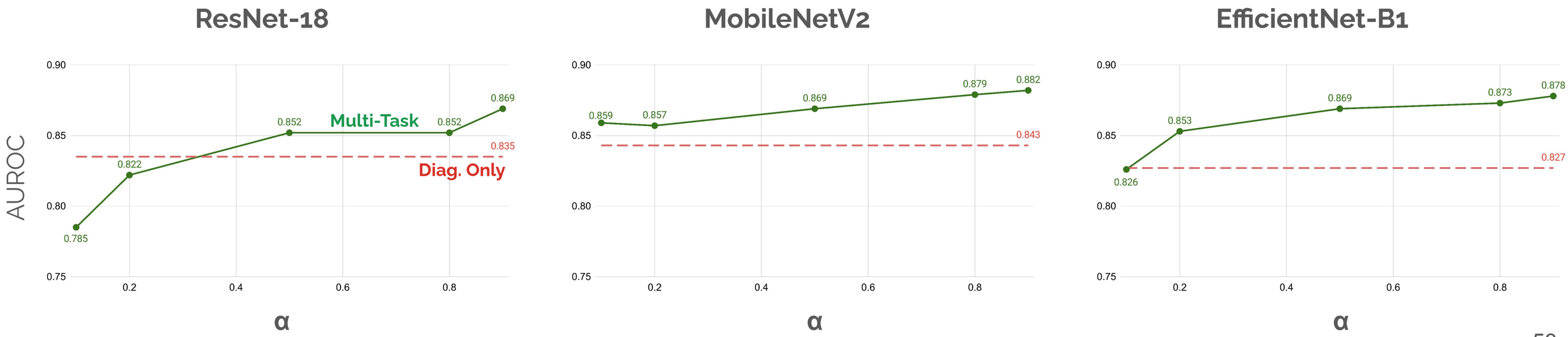


Multi-Task Models Diagnose Better than Diag. Only Models

Experiment: Vary α to study the relative importance of IAA prediction.

Results:

- Diagnosis-dominant (α = 0.9) multi-task models perform the best.
- Multi-task models outperform diagnosis-only models.

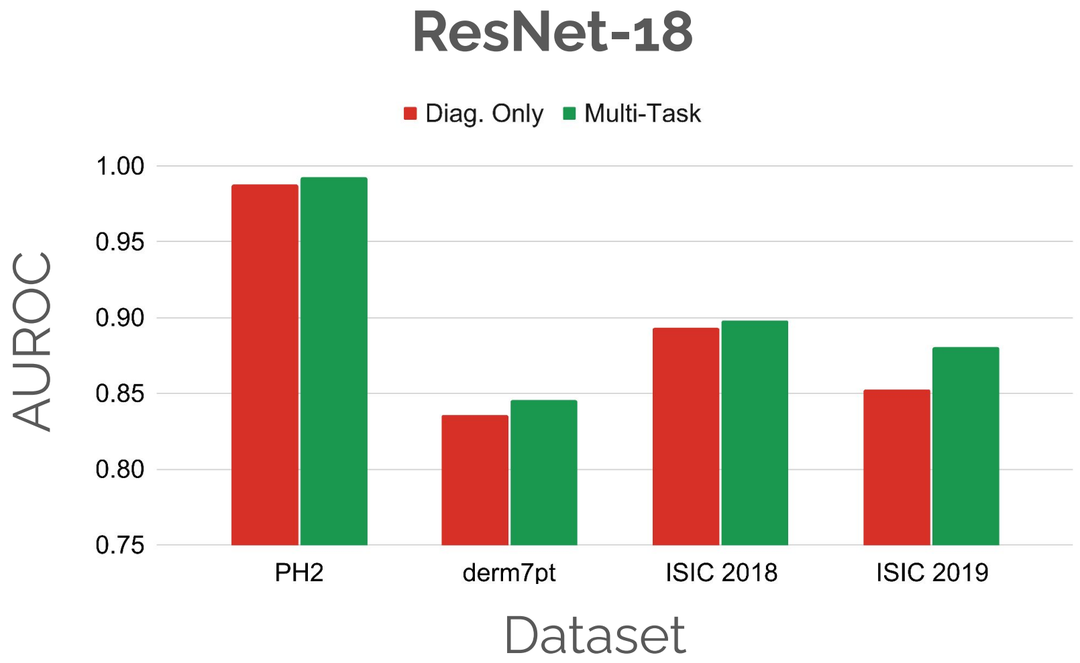


What about IAA-Aware Diagnosis on External Datasets?

Multi-task models, trained on IMA++, **fine-tuned on 4 dermoscopic datasets**.

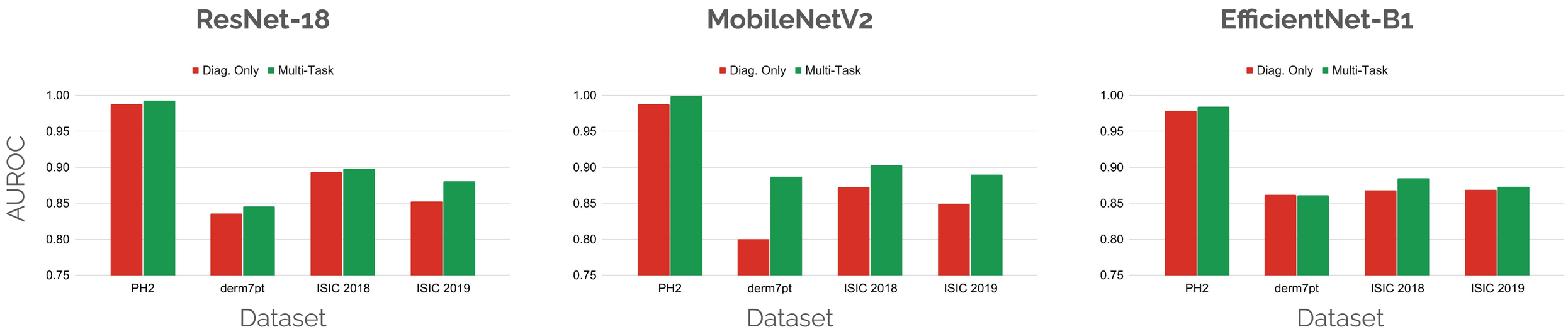
IAA-Aware Diagnosis Improves Performance on Other Datasets

Multi-task models, trained on IMA++, **fine-tuned on 4 dermoscopic datasets**.



IAA-Aware Diagnosis Improves Performance on Other Datasets

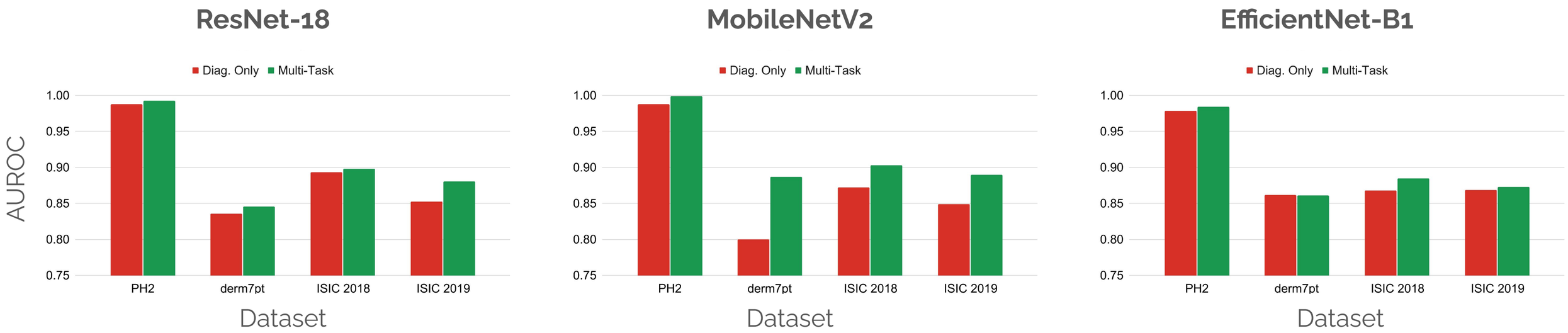
Multi-task models, trained on IMA++, **fine-tuned on 4 dermoscopic datasets**.



IAA-Aware Diagnosis Improves Performance on Other Datasets

Multi-task models, trained on IMA++, **fine-tuned on 4 dermoscopic datasets**.

Performance gains may be transferable: Collect multi-annotator masks once (IMA++), transfer gains downstream to **single-annotator datasets**.



Conclusion

- IMA++ enables the largest skin lesion segmentation variability study.

Conclusion

- IMA++ enables the largest skin lesion segmentation variability study.

- Benign lesions show higher IAA than malignant (distribution shift).

Conclusion

- IMA++ enables the largest skin lesion segmentation variability study.
- Benign lesions show higher IAA than malignant (distribution shift).
- Predicting IAA and diagnosis in a multi-task framework improves diagnostic performance, including on single-annotator datasets.

Conclusion

- IMA++ enables the largest skin lesion segmentation variability study.
- Benign lesions show higher IAA than malignant (distribution shift).
- Predicting IAA and diagnosis in a multi-task framework improves diagnostic performance, including on single-annotator datasets.
- **Future work:** Is averaging a pairwise metric (Dice, Hausdorff distance) the best way to capture groupwise IAA?

References

[1] Leung et al., “Pulmonary nodule - Benign versus Malignant: Differentiation with CT and PET-CT”, https://radiologyassistant.nl/chest/solitary-pulmonary-nodule/benign-versus-malignant, 2007.

[2] MacMahon et al, “Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017”, *Radiology*, 2017.

[3] Glassman et al., “MRI of the Breast”, https://radiologyassistant.nl/breast/mri/mri-of-the-breast, 2009.

[4] Kittler et al., “Chaos and Clues”, https://dermoscopedia.org/Chaos_and_Clues.

[5] Menzies et al., “The morphologic criteria of the pseudopod in surface microscopy”, *Archives of Dermatology*, 1995.

[6] Williams et al., “Assessment of Diagnostic Accuracy of Dermoscopic Structures and Patterns Used in Melanoma Detection”, *JAMA Dermatology*, 2021.

Thank you.

Questions?

kabhishe@sfu.ca

https://github.com/sfu-mial/skin-IAV

Acknowledgements

NSERC CRSNG

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

NVIDIA