Biomarker estimation from medical images: segmentation-based and segmentation-free approaches

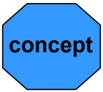
Jing ZHANG (张 晶)

LITIS, INSA de Rouen, 06-04-2022

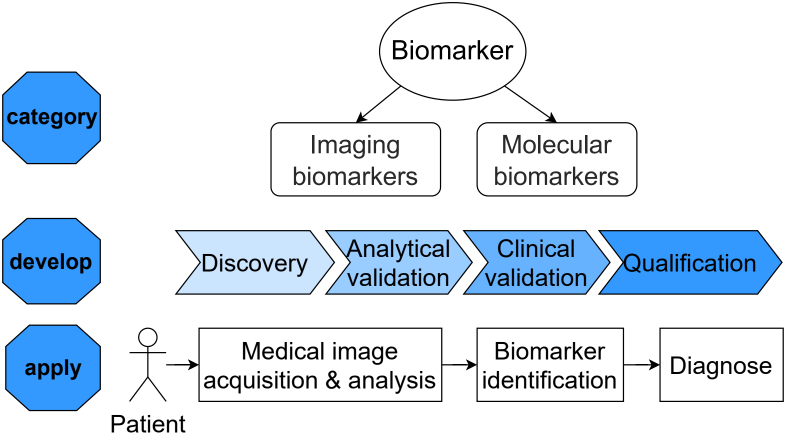
Jury members

M. Désiré SIDIBÉ	PR de l' Université Evry Val Essonne	Rapporteur
M. Aymeric HISTACE	PR de l' ENSEA, Cergy	Rapporteur
Mme Mireille GARREAU	PR de l' Université de Rennes I	Examinatrice
Mme Caroline PETITJEAN	PR de l' Université de Rouen Normandie	Codirectrice de thèse
Mme Samia AINOUZ	PR de l' INSA Rouen Normandie	Directrice de thèse

Biomarker



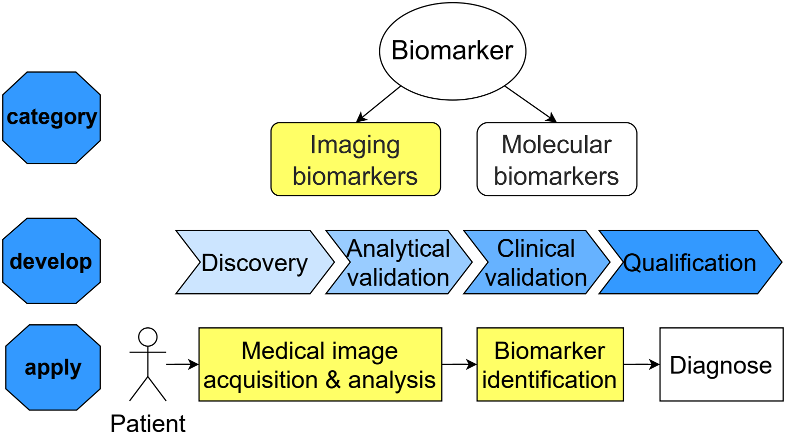
A biomarker is anything that can be used as an indicator of a particular disease state.



Biomarker

concept

A biomarker is anything that can be used as an indicator of a particular disease state.

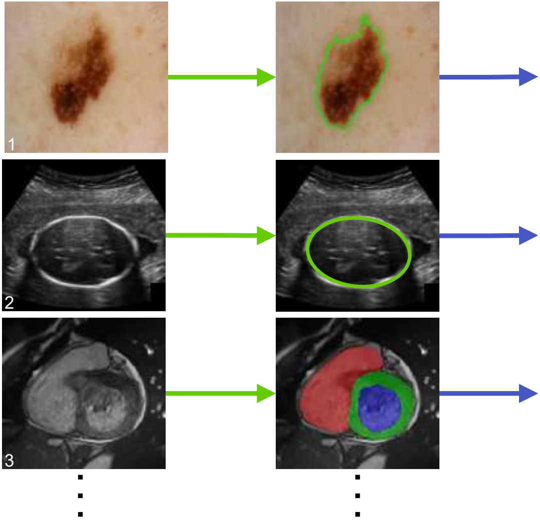


Jing ZHANG
LITIS,
INSA de Rouen

Generally...

How to measure a **biomarker**?

Biomarker identification examples



Skin cancer indices:
Asymmetry
Border
Color
Diameter
Evolving

Head **C**ircumference

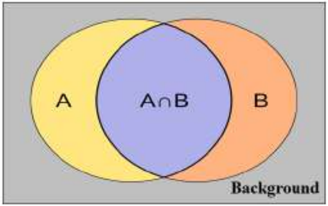
Cardiac structure volume

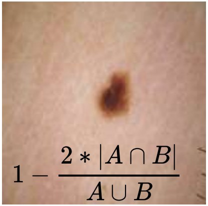
Image credit: 1. `https://challenge2018.isic-archive.com/`; 2. `https://hc18.grand-challenge.org/`; 3. `https://acdc.creatis.insa-lyon.fr/`

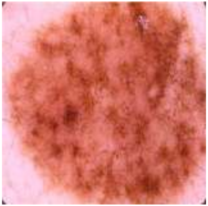
Problems

Medical image → **Segmentation** → **Biomarker identification**

Segmentation: ☹ Intermediate step, prone to errors

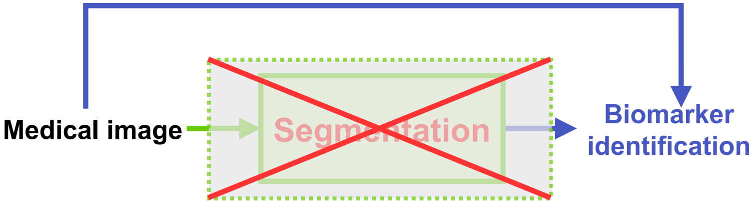






Dice loss in Deep learning: ☹ Does not generalize well

Very recently...



Direct estimation methods:

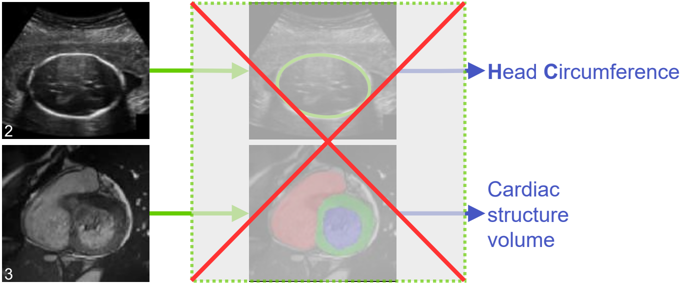
- ☺ Without segmentation intervention
- ☹ Less explainable

Contributions

Biomarker estimation:

- Segmentation-based
  - ① **New Kappa loss function**
- Segmentation-free (novel)
  - ② **Direct** biomarker estimation **without segmentation**
  - ③ **Explainability** of segmentation-free models

VS



Thesis outline

2. Kappa loss for skin lesion segmentation [Chapter 3]

3. Direct fetus head circumference estimation [Chapter 4]
   - HC estimation via Seg vs. Seg-free
   - Explainability for regression CNN

4. Direct cardiac structure volume prediction [Chapter 5]

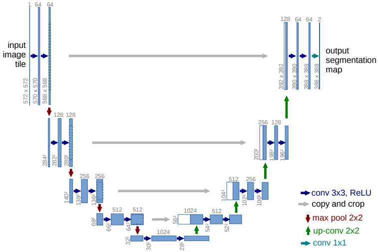
5. Conclusions and Perspectives [Chapter 6]

Kappa loss for skin lesion segmentation in fully convolutional network

Background

Deep segmentation models

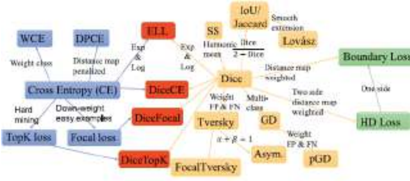
- **U-Net**[1]
- FCN[2]
- V-Net[3] ...



U-Net architecture[2]

Loss functions

- Region-based: **Dice loss**[3]
- Cross entropy[4]
- Distance-based loss
- Hybrid loss

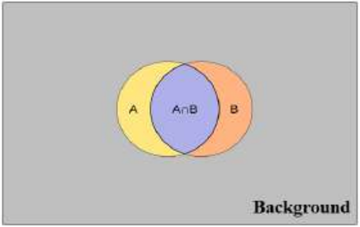


Losses overview[5]

1.Ronneberger et al., 2015; 2.Long et al., 2015; 3.Milletari et al., 2016; 4.Cox, 1959; 5.Ma et al., 2021

Dice loss

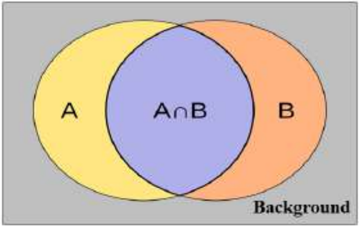
Class imbalance problem

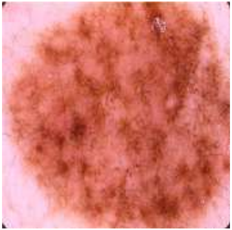


$$\text{Dice metric} = \frac{2 * |A \cap B|}{A \cup B} \tag{1}$$

Dice loss

Problem: When classes are balanced





$$\text{Dice loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N}(p_i + g_i)} \tag{2}$$

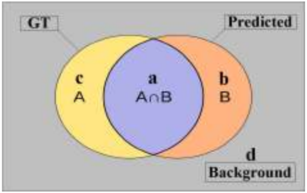
☹ Background is discarded in Dice loss!

p_i: probability of predicted pixels;
g_i: ground truth pixels.

Kappa index

❑ **Kappa index** (Cohen, 1960): the agreement between two raters.

		GT (Rater 1)		
		+	-	Total
Predicted	+	a	b	$a+b$
results	-	c	d	$c+d$
(Rater 2)	Total	$a+c$	$b+d$	N



$$\kappa = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{3}$$

When $d \gg (a; b; c)$, Eq (3) becomes:

$$\lim_{d \to \infty} \kappa = \frac{2a}{2a+b+c} \to \text{Dice}$$

Proposed Kappa loss

❑ **Proposed Kappa loss**:

$$\text{Kappa loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i - \sum_{i=1}^{N} p_i \cdot \sum_{i=1}^{N} g_i / N}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i - 2\sum_{i=1}^{N} p_i \cdot \sum_{i=1}^{N} g_i / N} \tag{4}$$

- ✘ Dice loss ignores the background pixels.
- ✔ Kappa loss considers all pixels.
- ✔ Kappa loss is generalization of Dice loss.

p_i: probability of predicted pixels;
g_i: ground truth pixels;
N: the number of pixels.

Experimental results

- **Datasets**: Skin lesion (6), data augmentation
- **Segmentation Model**: U-Net, train from scratch
- **Loss functions**: Dice loss; Kappa loss
- **Evaluation metrics**: Dice index ↑; Hausdorff distance ↓

dataset(#)	Dice loss		Kappa loss	
	DI ↑	HD ↓	DI ↑	HD ↓
Non-mel(87)†	0.65±0.11	5.06±1.79	**0.73±0.11**	**4.70±2.02**
Mel(119)†	0.80±0.06	6.70±1.93	0.81±0.03	6.59±1.88
SCD(206)†	0.82±0.04	7.94±1.72	0.83±0.03	7.91±1.68
ISIC-16(900)❖	0.80±0.05	8.42±2.19	**0.84±0.01**	8.41±2.25
ISIC-17(2000)❖	0.80±0.05	8.07±1.93	**0.84±0.05**	8.03±1.94
ISIC-18(2594)❖	0.81±0.03	7.59±2.60	0.82±0.04	7.52±2.66

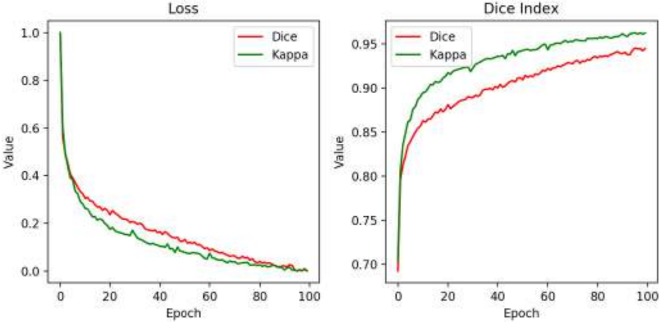
✔ Kappa loss > Dice loss

- Input data & scale

† Skin-Cancer-Detection (SCD, Non-mel, Mel):
https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection

❖ International Skin Imaging Collaboration (ISIC):
https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

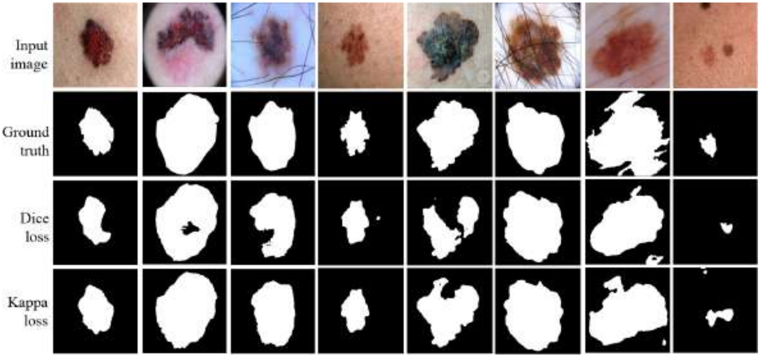
Learning curve



Loss curve (left) and Dice index (right) of training process:

- **Convergence**: Kappa loss (green) > Dice loss (red);
- **Dice index**: Kappa loss (green) > Dice loss (red).

Qualitative results



✘ Dice loss: Defects/holes in the skin lesion.
✔ Kappa loss: Less segmentation error.

Summary of Kappa loss

- Takes into account all the pixels → more accurate.
- Kappa loss is generalization of Dice loss.
- Performs as good as Dice loss → background information.
- Segmentation results as prerequisite for biomarker identification → direct estimation.

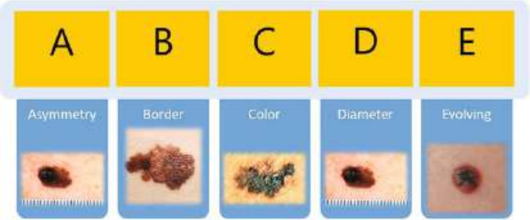


Image credit: https://doi.org/10.1007/s11063-020-10364-y

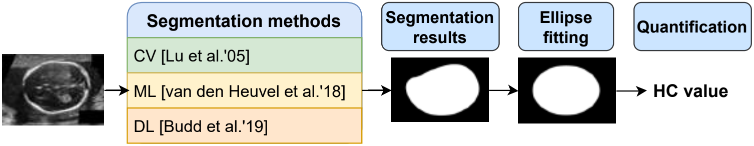
IEEE ISBI 2020 International Symposium on Biomedical Imaging

https://github.com/jizhang02/Kappa-loss

Direct fetus head circumference (HC) estimation from ultrasound images

Related works

➤ HC measurements → **Segmentation**(State-of-the-art):



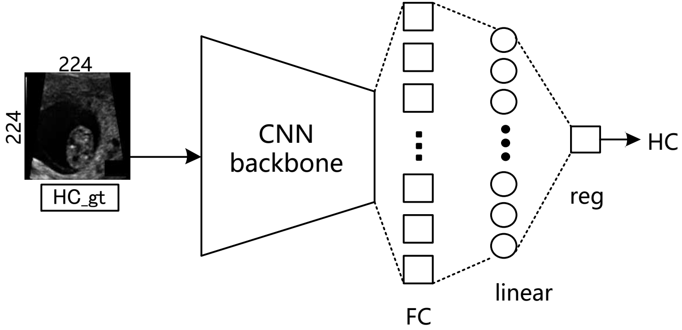
✔ Pros: Have segmentation results, **explainable**.

✘ Cons: Multi steps(Seg+ellipse fitting), prone to errors.

➤ Direct measurements → **Regression** methods:

- Head pose, facial landmark (Riegler et al., 2013)
- Left ventricular volumes prediction (Luo et al., 2016)
- Kidney volume prediction (Hussain et al., 2016)
- Brain age (Li et al., 2018)

Proposed Regression CNN



Benefits of proposed method:

- Doesn't need ground truth images, no segmentation errors.
- Can estimate the HC value directly.

Proposed regression CNN

2 changes from classification CNN to regression CNN model:

- ~~Softmax layer~~ → Linear regression layer
- ~~Classification loss~~ → Regression loss
  - $MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - g_i|$
  - $MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - g_i)^2$
  - $HL^* = \begin{cases} \frac{1}{n}\sum\limits_{i=1}^{n}\frac{1}{2}(p_i - g_i)^2, & \text{for} \quad |p_i - g_i| < \delta \\ \frac{1}{n}\sum\limits_{i=1}^{n}\delta * (|p_i - g_i| - \frac{\delta}{2}), & \text{otherwise} \end{cases}$

*HL: Huber Loss;
p_i = predicted values;
g_i = ground truth values

Experiment protocol

Models

- Segmentation models+CNN backbones, Dice loss
- Regression models+CNN backbones, MAE, MSE, HL loss
- Transfer learning: pretrained on ImageNet[†], fine-tuned on HC18.

Settings

- HC18[‡] dataset, 5-fold cross validation
- Preprocessing: Data augmentation, Normalization, Resizing
- Metrics: MAE, percentage of MAE (PMAE)

[†] https://www.image-net.org/

[‡] https://hc18.grand-challenge.org/

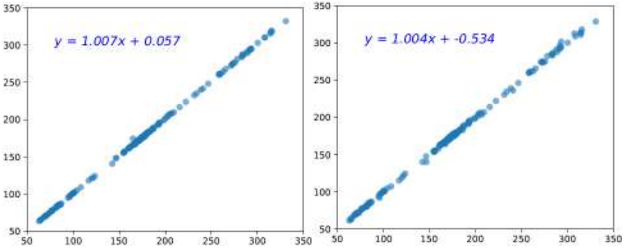
Model performance: Seg-based vs. Seg-free

Metrics	**MAE(mm)**	**MAE(px)**	**PMAE(%)**
Methods	Segmentation-based methods		
U-Net-B2	**1.08±1.25**	**7.87±7.51**	**0.65±0.68**
LinkNet-B2	1.15±1.32	8.45±8.39	0.69±0.77
	Segmentation-free methods (Our methods)		
Reg-B3-L1	**1.83±2.11**	**13.57±13.53**	**1.17±1.43**
Reg-B3-L2	2.35±2.74	17.32 ±17.95	1.53 ±2.02

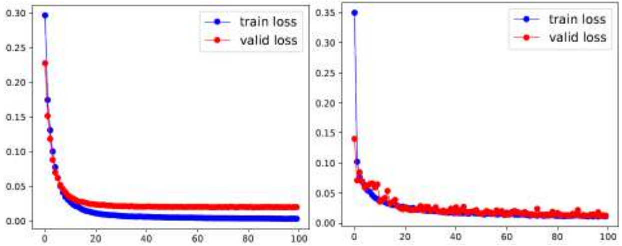
- Prediction error:
  - ☹ Seg-free > Seg-based methods
  - 😐 Same order of magnitude
  - ☺ Seg-based & Seg-free methods $\ll$ inter-operator variability (±12 mm, Sarris et al., 2012).

B2: ResNet50; B3: EfficientNet; L1: MAE loss; L2: MSE loss

Qualitative analysis: Seg-based vs. Seg-free



(a) U-Net-B2 (b) Regression-B3



(c) U-Net-B2 (d) Regression-B3

B2: ResNet50; B3: EfficientNet

Computation efficiency: Seg-based vs. Seg-free

Methods	Train (s/epoch)	Predict (s/test set)	Mem-M (GB)	Mem-P (GB)
Segmentation-based methods				
U-Net-B2	29	68.26	3.06	1.84
LinkNet-B2	30	80.36	3.82	1.90
Segmentation-free methods (Our methods)				
Reg-B2-L1	20	48.28	2.31	1.73
Reg-B3-L1	38	36.95	2.29	2.68

- ☺ **Time**: Seg-based > Seg-free methods
- ☺ **Memory**: Seg-based > Seg-free methods

B2: ResNet50; B3: EfficientNet; L1: MAE loss; M: Model memory; P: Practical memory

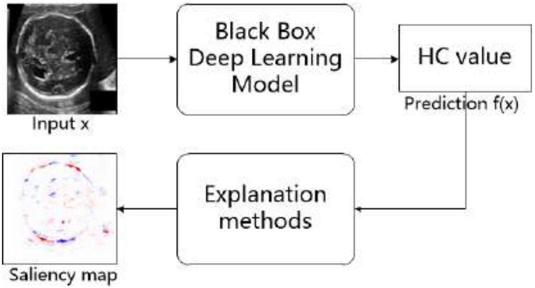
Summary of HC estimation

- Proposing **regression CNN model** → directly estimate the HC value without segmentation intervention.
- Evaluating Seg-based vs. Seg-free methods fairly.
- Trade off between prediction error and computation efficiency.
- Limitation: Less explainable.

MIDL Montréal 2020 [Journal of Imaging'22]

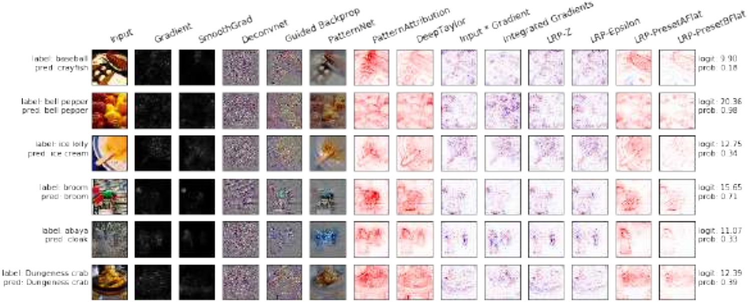
https://github.com/jizhang02/HC-reg-seg

Explainability for regression CNN



Saliency maps: Highlight features that model learns from images

- Verify prediction results
- Find out flaws and biases



Saliency maps of classification models†.

Explanation methods:

Gradient

- Gradient[1]
- SmoothGrad[2]
- Input*Gradient[3]

Deconvolution

- Integrated Gradients[4]
- DeConvNet[5]
- Guided BackProp[6]

LRP var.

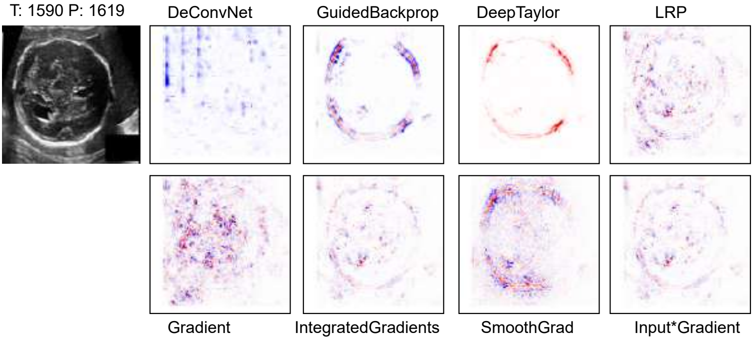
- LRP[7]
- DeepTayor[8]

☞ Which method is suitable?

†Image credit: `https://github.com/albermax/innvestigate`

1.Simonyan et al., 2014; 2.Smilkov et al., 2017; 3.Shrikumar et al., 2016; 4.Sundararajan et al., 2017; 5.Zeiler et al., 2014; 6.Springenberg et al., 2015; 7. Layer-wise Relevance Propagation, Back et al., 2015; 8.Montavon et al., 2017

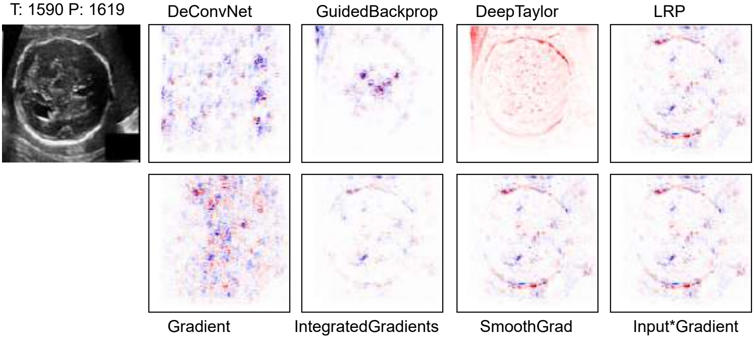
Saliency maps on regression CNN



Saliency maps on regression VGG16

☹ No effective features are retrieved by DeConvNet and Gradient.
P:predicted HC value, T: ground truth HC value (in pixels).

Saliency maps on regression CNN



Saliency maps on regression ResNet50

☹ DeConvNet and Gradient are insensitive to the models.

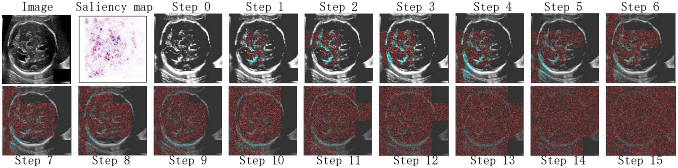
☹ Quantitative criteria is needed.

P:predicted HC value, T: ground truth HC value (in pixels).

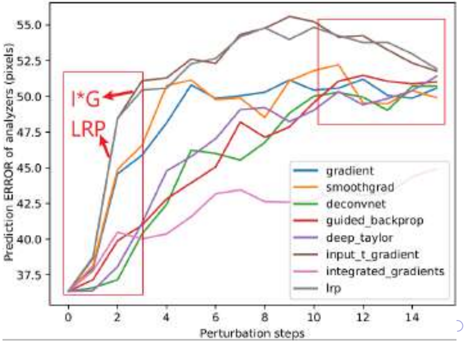
Quantitative evaluation method

Perturbation-based evaluation:

➥ Adding noise according to **importance rank** of saliency map



Perturbation process



Quantitative evaluation method

Area over Perturbation Curve (AOPC) (Samek et al., 2017)

➥ Difference of accuracy with/without perturbation in **Classification**

Adapted for Regression (Accuracy → Error $\epsilon(x_n)$)

$$\mathrm{AOPC}_{Analyzer}^{regression} = \frac{1}{N}\sum_{n=0}^{N}\left(\epsilon(x_n)^{(0)} - \frac{1}{K}\sum_{k=0}^{K}\epsilon(x_n)^{(k)}\right) \quad (5)$$

N is the number of images, K is the number of perturbation steps, ϵ is prediction error.

AOPC score:

- ↑ Models: Sensitive to perturbation
- ↑ Explanation method: Ability to capture highlight features

Quantitative evaluation result

AOPC score of different explanation methods after perturbation

Model	G	SG	DCN	DT	GB	I*G	IG	LRP
Reg_VGG	-7.31	-7.39	-2.87	-7.40	-1.66	**-9.19**	**-9.49**	**-9.17**
Reg_ResN	-11.53	-11.84	-9.25	-9.89	-9.72	**-14.75**	-5.60	**-14.58**

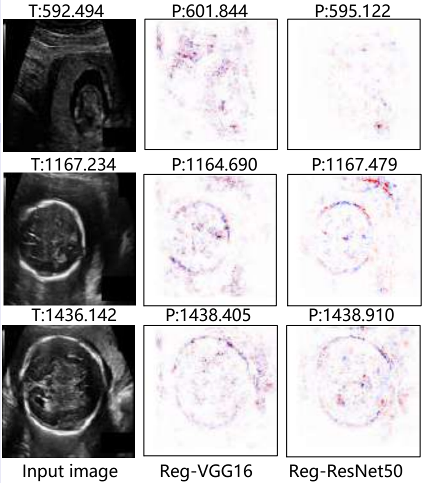
Note: G: Gradient, SG: SmoothGrad, DCN: DeConvNet, DT: DeepTaylor, GB: GuidedBackprop, I*G: Input*Gradient, IG: IntegratedGradients. Lower is better. Best scores in bold.

AOPC score:

- ☺ I*G and LRP explanation methods are better
- ☺ Regression ResNet > Regression VGG

Model's explainability Validation

Saliency maps for **good prediction results** on two models

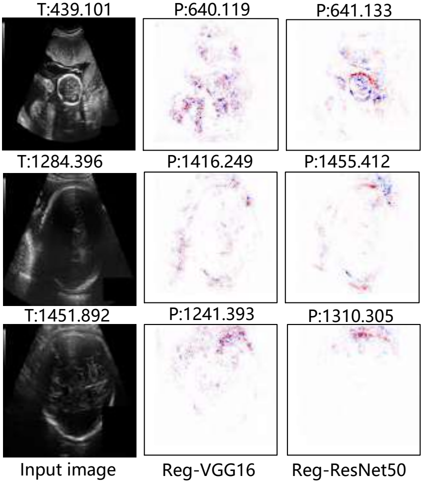


P and T: predicted and ground truth HC values (pixels).

- Input*Gradient method
- Prediction results: **Explainable**
- Reg-ResNet > Reg-VGG

Model's explainability Validation

Saliency maps for **bad prediction results** on two models



P and T: predicted and ground truth HC values (pixels).

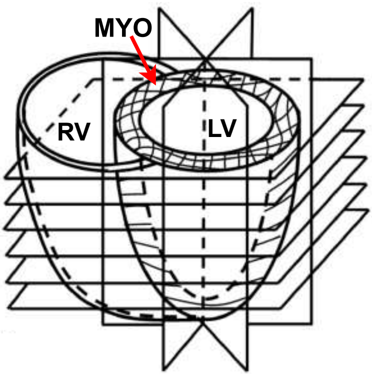
- Input*Gradient method
- Flaws and biases

Summary of Explainability

- Using **post-hoc explanation techniques** to valid the explainability of **regression CNN**.
- Evaluating explanation methods through **saliency maps** and **perturbation based methods**.
- Extending model type from **classification** to **regression**.

https://github.com/jizhang02/XAI-reg

Direct cardiac multi-structure volume prediction from 3D MRI images



Cardiac structure

RV: Right Ventricle

MYO: Myocardium

LV: Left Ventricle

ESV: End Systole Volume

EDV: End Diastole Volume

$$\text{Ejection fraction} = \frac{\text{EDV}-\text{ESV}}{\text{EDV}} * 100$$

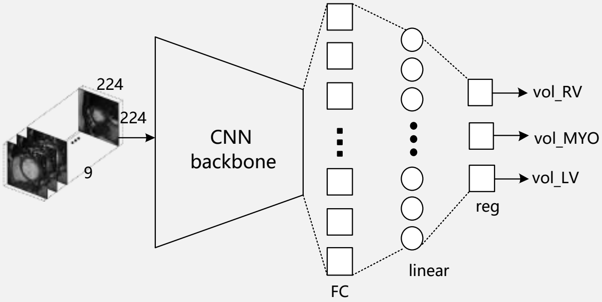
Image credit: `https://doi.org/10.1186/1532-429X-13-36`

Methodology

Goal

Predict the volume of cardiac structures directly without segmentation.

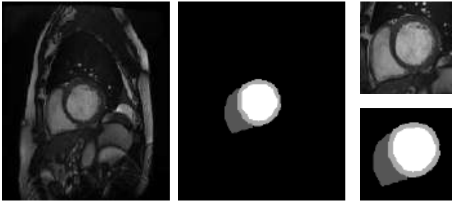
Solution–Regression CNN



ACDC dataset*

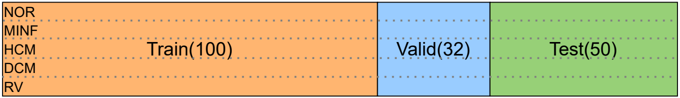
Preprocessing

- Slice unifying (9)
- Slice cropping (100*100)
- Data augmentation



P001_fm01 P001_fm01_gt crpped_P001_fm01

➤ Dataset: 182 subjects



➤ Ground truth: volume of RV, MYO, LV (ml)

$$\text{Volume} = \sum_{i}^{N} \mathrm{S_i} * \mathrm{px\_x} * \mathrm{px\_y} * \mathrm{px\_z}/1000 \quad (6)$$

S: Area of cardiac structure; *N*: Number of cardiac slices; *px*: pixelsize in x,y,z

∗ `https://acdc.creatis.insa-lyon.fr/`

Experimental results

- **Data**: MRI cardiac
- **Model**: ① Reg_VGG;
  ② Reg_ResNet;
  ③ Reg_efn*
- **Results**: 5-fold cross validation
- **Loss**: MAE loss
- **Metrics**: MAE ↓; PMAE ↓

Prediction error of 3 cardiac structures on regression CNNs:

	RV		MYO		LV	
	MAE(ml)	PMAE(%)	MAE(ml)	PMAE(%)	MAE(ml)	PMAE(%)
①	50.51±39.81	65.64±92.96	41.58±34.38	36.43±37.59	35.29±29.49	40.20±52.63
②	**43.11±36.57**	**51.63±69.46**	**36.98±29.25**	**31.96±29.44**	**33.19±26.48**	**39.09±47.45**
③	49.55±40.88	60.76±83.83	36.70±32.26	33.50±36.88	33.51±26.93	39.78±52.15

➤ Reg_ResNet > Reg_ VGG, Reg_efn

➤ RV is difficult to predict

* EfficientNet, Tan et al. 2019

Experimental results

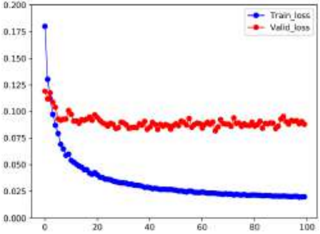
Prediction error of LV and comparison to SotA:

	MAE(ml)	PMAE(%)
	Segmentation-free	
Reg_ResNet	33.19±26.48	39.09±47.45
Luo (2020a)	7.55±4.5	—
Luo (2020b)	5.1±3.2	—
Zhen (2016)	12.5±8.4	—
	Segmentation-based	
Zheng (2018)	12±9.1	—
Vigneault (2018)	11.1±5.3	—
Liao (2017)	12.85±9.55	—
Ngo (2017)	16.95±12	—
Avendi (2017)	18.35±18.6	—

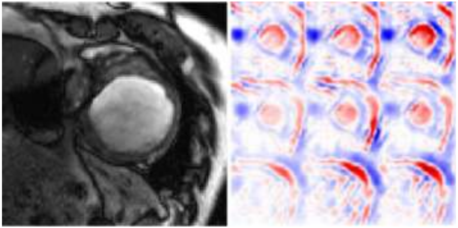
☹ Improvement room is left

☹ Results are not so comparable

Discussion



(a) Loss curve



(b) Saliency map

- Regression CNN model, **over-fitting**
- **Low SNR** in cardiac images.
- The **cardiac slices** may influence the model's prediction

Summary

- Estimating cardiac structure volume directly.
- Extending image modality and prediction targets.
- Experimental results have room for improvement.

Conclusions of thesis

New loss:

- ☑ Kappa loss
- ☑ Generalization of Dice loss
- ☑ Class imbalance & balance problem

Biomarker estimation:

- ☑ Regression CNN—Segmentation-free
- ☑ Head circumference: Seg vs. Seg-free
- ☑ Cardiac structure volume

Model's explainability:

- ☑ Saliency maps on regression CNNs
- ☑ Evaluate explanation methods: Perturbation-based

Perspectives

Kappa loss—Segmentation

- Verify on more medical datasets
- Split dataset according to size of target

Biomarker estimation: Regression CNN—Segmentation-free

- Verify other biomarkers
- Multi-task learning
- Attention mechanisms
- Geometric deep learning

Model's explainability:

- Self explainable
- Applicable in clinical application

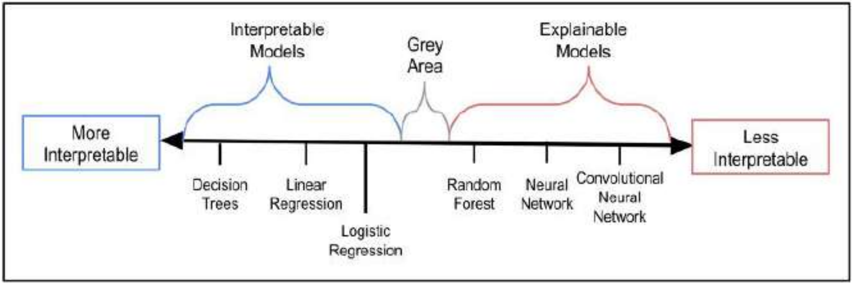
The end

Thank you for your attention!

Acknowledgment:

- € China Scholarship Council (CSC)
- Centre Régional Informatique et d'Applications Numériques de Normandie (CRIANN)

Interpretable and Explainable



Ref: https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

Additional info

AOPC score of different losses:

Table A.1 – Performance (AOPC scores) of different explanation methods after perturbation, with two regression models and three loss functions. G: Gradient, SG: SmoothGrad, DCN: DeConvNet, DT: DeepTaylor, GB: GuidedBackprop, I*G: Input*Gradient, IG: IntegratedGradients. Lower is better. Best scores in bold.

Model	G	SG	DCN	DT	GB	I*G	IG	LRP
RegVGG_MAE	-7.31	-7.39	-2.87	-7.40	-1.66	**-9.19**	**-9.49**	**-9.17**
RegVGG_MSE	-7.80	-7.18	-5.36	-9.10	-2.99	**-14.57**	N/A	**-14.46**
RegVGG_HL	-23.39	-21.63	-24.59	-27.78	-18.86	**-29.47**	N/A	**-29.27**
RegResNet_MAE	-11.53	-11.84	-9.25	-9.89	-9.72	**-14.75**	-5.60	**-14.58**
RegResNet_MSE	-11.31	N/A	-11.18	-19.41	N/A	**-32.48**	**-20.49**	**-32.51**
RegResNet_HL	-24.17	-24.27	N/A	-22.66	-28.42	**-37.12**	-22.81	**-38.12**

Additional info

AOPC score of different losses on cardiac structure volume:

Table 5.4 – Prediction error on volume of 3 cardiac structures using regression CNN models with 3 different loss functions, MAE loss, MSE loss and Huber loss (HL). The 3 CNN backbones are VGG16, ResNetV2, EfficientNetb2 (efn). The models are trained on 2900 training images. The results are average results of 5-fold cross validation.

Model	MAE_RV(ml)	PMAE(%)	MAE_MYO(ml)	PMAE(%)	MAE_LV(ml)	PMAE(%)
MAE loss						
Reg_VGG	50.51±39.81	65.64±92.96	41.58±34.38	36.43±37.59	35.29±29.49	40.20±52.63
Reg_ResNet	43.11±36.57	51.63±69.46	36.98±29.25	31.96±29.44	33.19±26.48	39.09±47.45
Reg_efn	49.55±40.88	60.76±83.83	36.70±32.26	33.50±36.88	33.51±26.93	39.78±52.15
MSE loss						
Reg_VGG	50.65±40.57	67.29±94.96	42.40±36.86	38.05±41.40	36.02±28.39	42.90±57.95
Reg_ResNet	43.82±34.47	56.33±79.52	38.11±29.74	34.63±33.99	33.28±26.27	42.01±53.51
Reg_efn	49.24±40.20	64.68±103.61	36.10±31.35	31.88±34.57	33.03±27.80	38.31±49.99
HL loss						
Reg_VGG	49.02±38.38	62.88±86.09	40.14±35.26	35.07±37.47	33.11±27.63	37.30±52.79
Reg_ResNet	49.29±38.12	66.20±93.30	35.13±31.27	30.99±32.11	32.75±28.30	41.57±56.69
Reg_efn	49.11±39.99	65.08±103.74	36.79±31.14	32.62±34.80	33.28±28.10	37.29±46.31

Additional info

Additional experiments on ACDC dataset:

- The influence of data modality
  - Cardiac slices
  - Training data scale
  - Single cardiac structure vs. Multi-structure
- Hyper parameters
  - Batchsize and learning rate
  - Dataset splitting
  - Data type distribution
- Vision Transformer

Additional info

Active contour loss for segmentation:

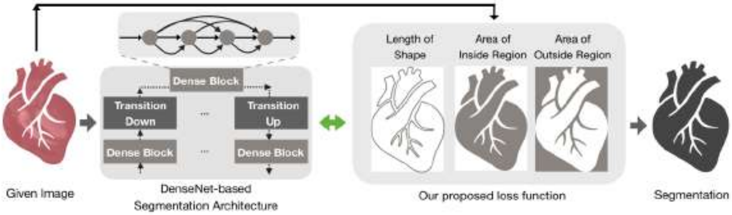


Figure 2: Overview of our proposed method which takes the object's area and length of the boundaries into account during training.

Chen, Xu, et al. "Learning active contour models for medical image segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Additional info

Robust loss for regression:

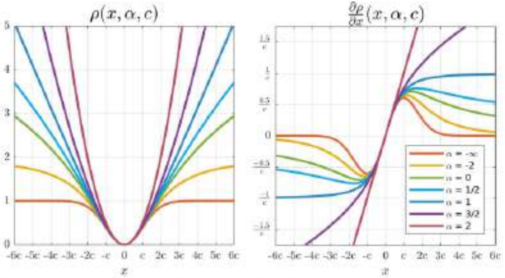


Figure 1. Our general loss function (left) and its gradient (right) for different values of its shape parameter α. Several values of α reproduce existing loss functions: L2 loss ($\alpha = 2$), Charbonnier loss ($\alpha = 1$), Cauchy loss ($\alpha = 0$), Geman-McClure loss ($\alpha = -2$), and Welsch loss ($\alpha = -\infty$).

Barron, Jonathan T. "A general and adaptive robust loss function." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.