Benchmarking Evaluation Metrics for Tubular Structure Segmentation in Biomedical Images

Meghane Decroocq, Charissa Poon, Matthias Schlachter, Henrik Skibbe

Brain Image Analysis Unit, Center for Brain Science, RIKEN, Wako, Japan

ShapeMI MICCAI Worshop, 27th Sept. 2025

Introduction

- **Tubular tree-like structures** are common in biology and medicine : vascular networks airways, neuronal trees

- The **segmentation** of these structures is crucial for many downstream applications : tracing, numerical simulation

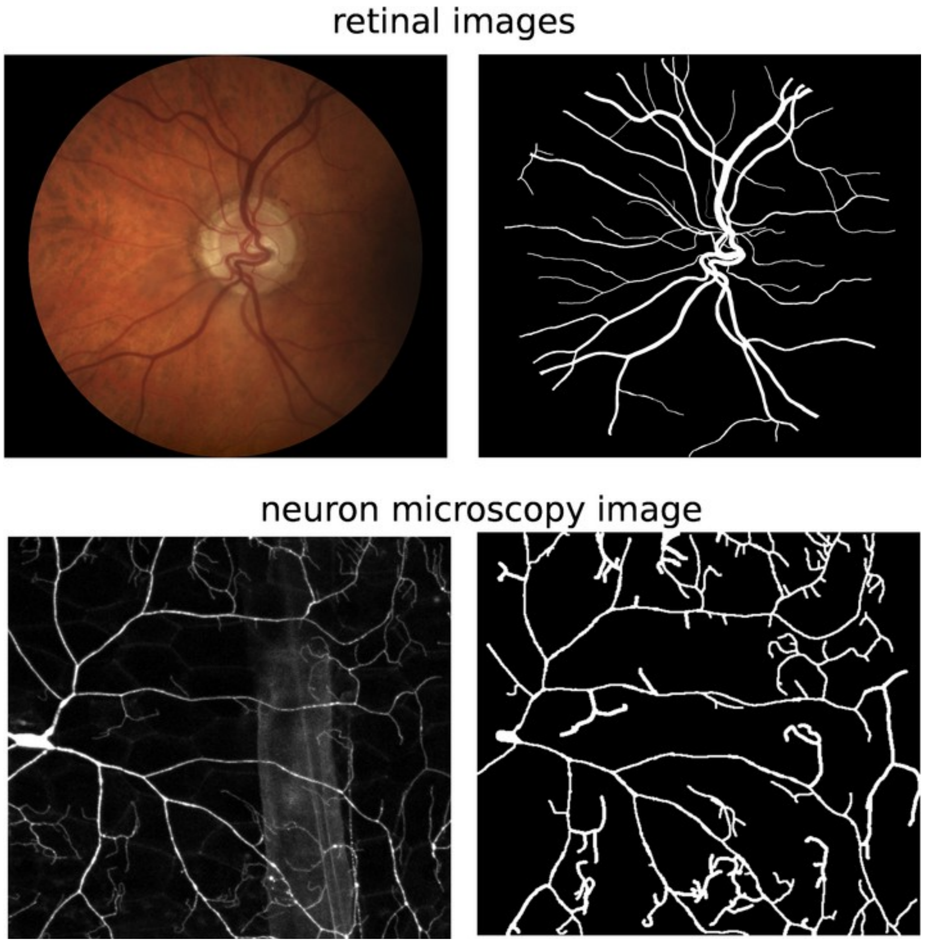


Figure 1. Example of biomedical images showing tubular, tree-like structures and the target segmentation.

Introduction

- **Deep learning** has led to tremendous progress in biomedical image segmentation.

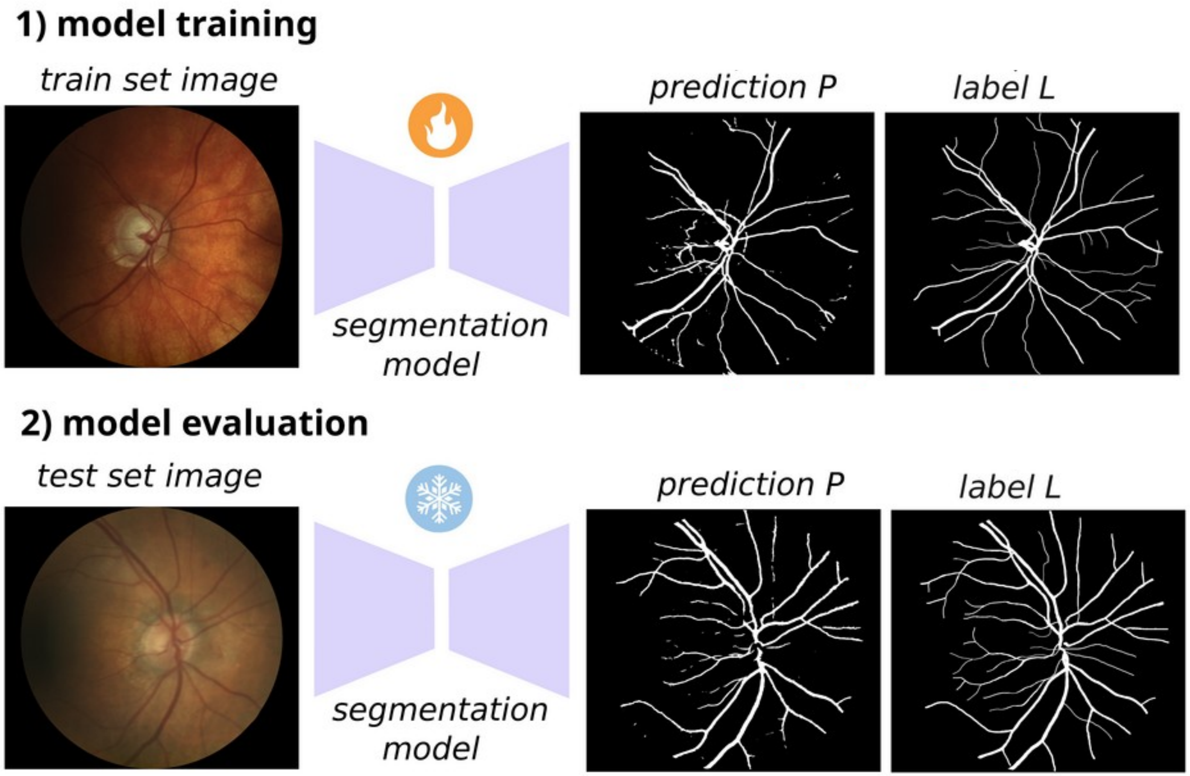


Figure 2. Importance of quality metrics.

Introduction

- To **train and evaluate** segmentation models, we need to design **metrics** measuring the **quality** of a predicted segmentation compared to the ground-truth labels.

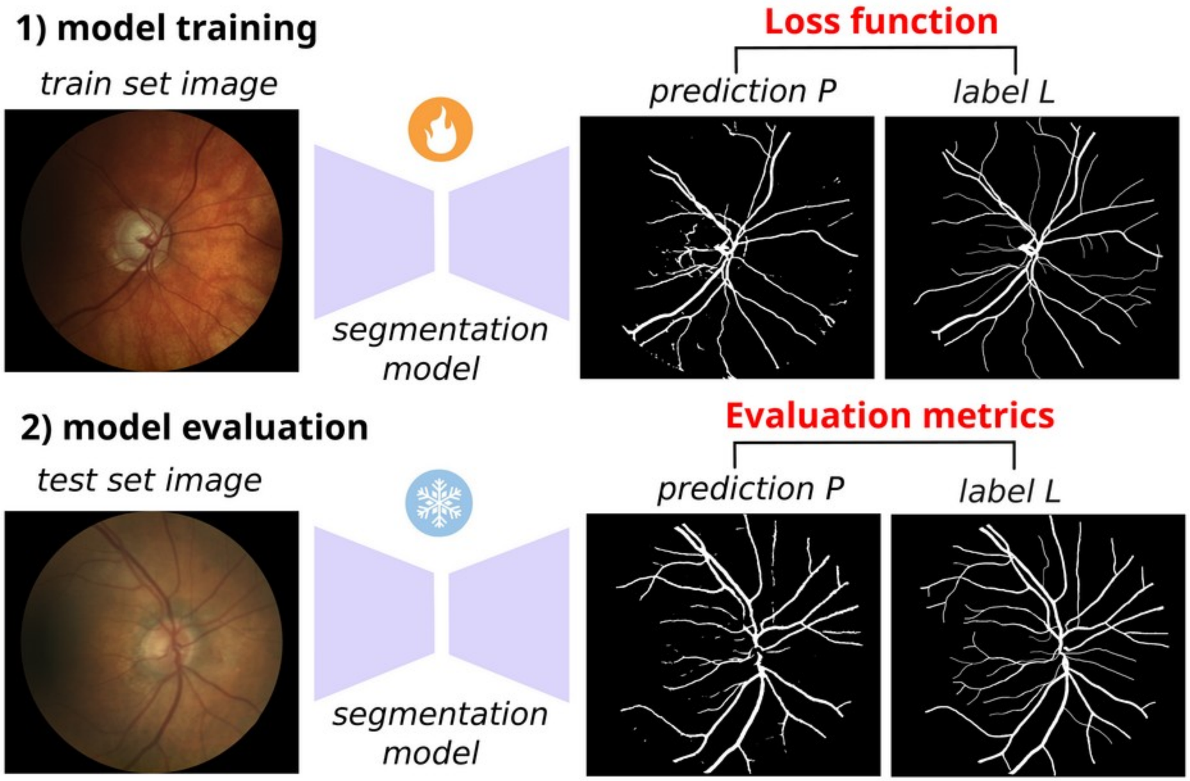


Figure 2. Importance of quality metrics.

Introduction

- Tubular structure segmentation offers unique **challenges**, leading to the design of **various metrics and loss functions** in recent years.

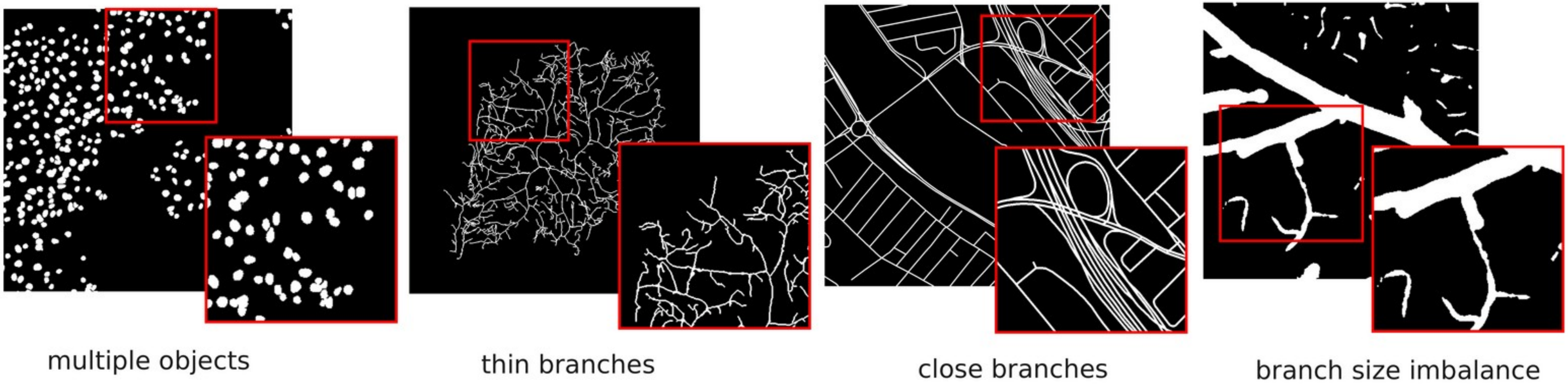


Figure 3. Challenges of tubular structure segmentation.

Introduction – Overlap-based quality metrics

- The **Dice** score is based on the pixel-wise mismatch between label and prediction.

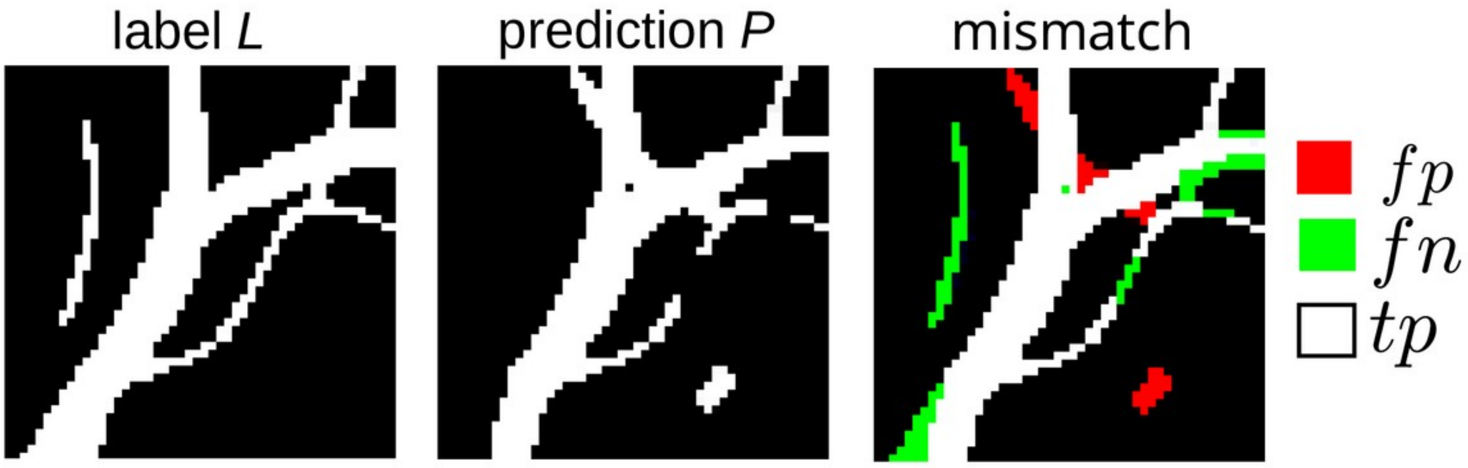


Figure 4. Illustration of the Dice calculation.
(tp = true positives, fn = false, fp = false postives)

Introduction – Skeleton-based quality metrics

- The **clDice**[1] relies on the **skeleton** mismatch.
- The **cbDice**[2] includes the **distance to boundaries.**

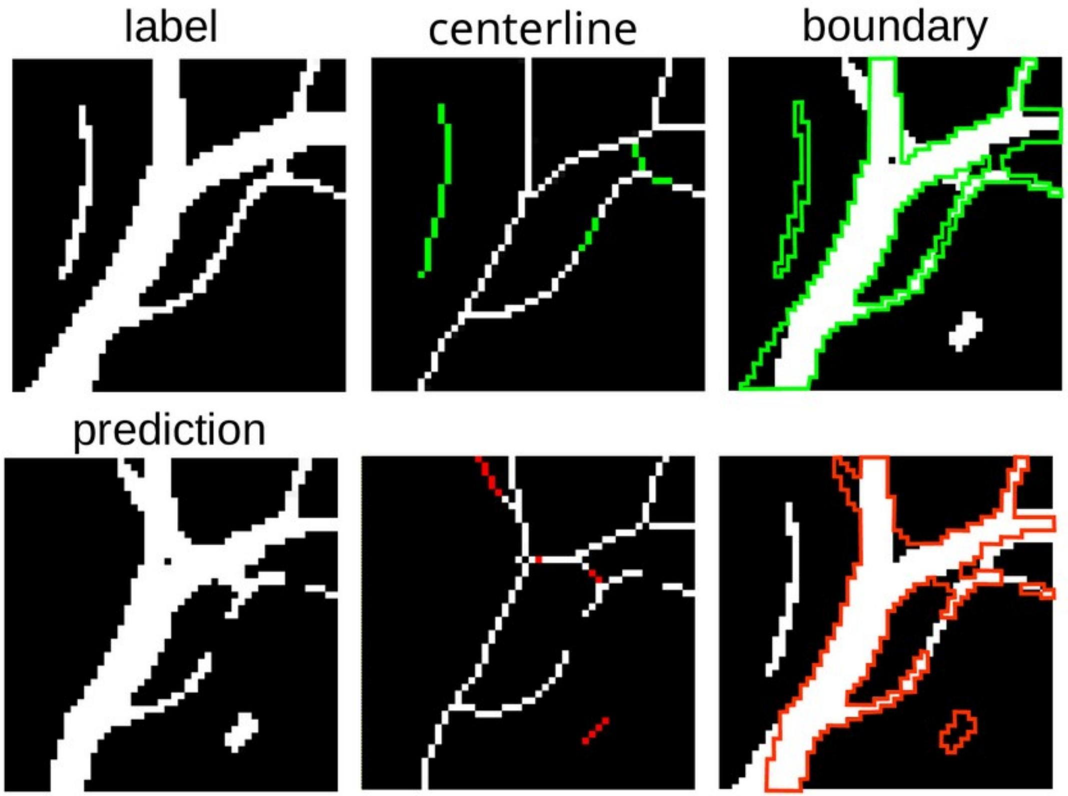


Figure 5. Illustration of the clDice and cbDice.

[1] Shit, Suprosanna, et al. "clDice-a novel topology-preserving loss function for tubular structure segmentation." IEEE/CVF, 2021.
[2] Shi, Pengcheng, et al. "Centerline boundary dice loss for vascular segmentation." MICCAI, 2024.

Introduction –
Algebraic topology quality metrics

β_0 = number of **connected components CC**
β_1 = number of **holes** in the image

- The **Betti error** β^{err} is the difference between the Betti numbers of the label and prediction.

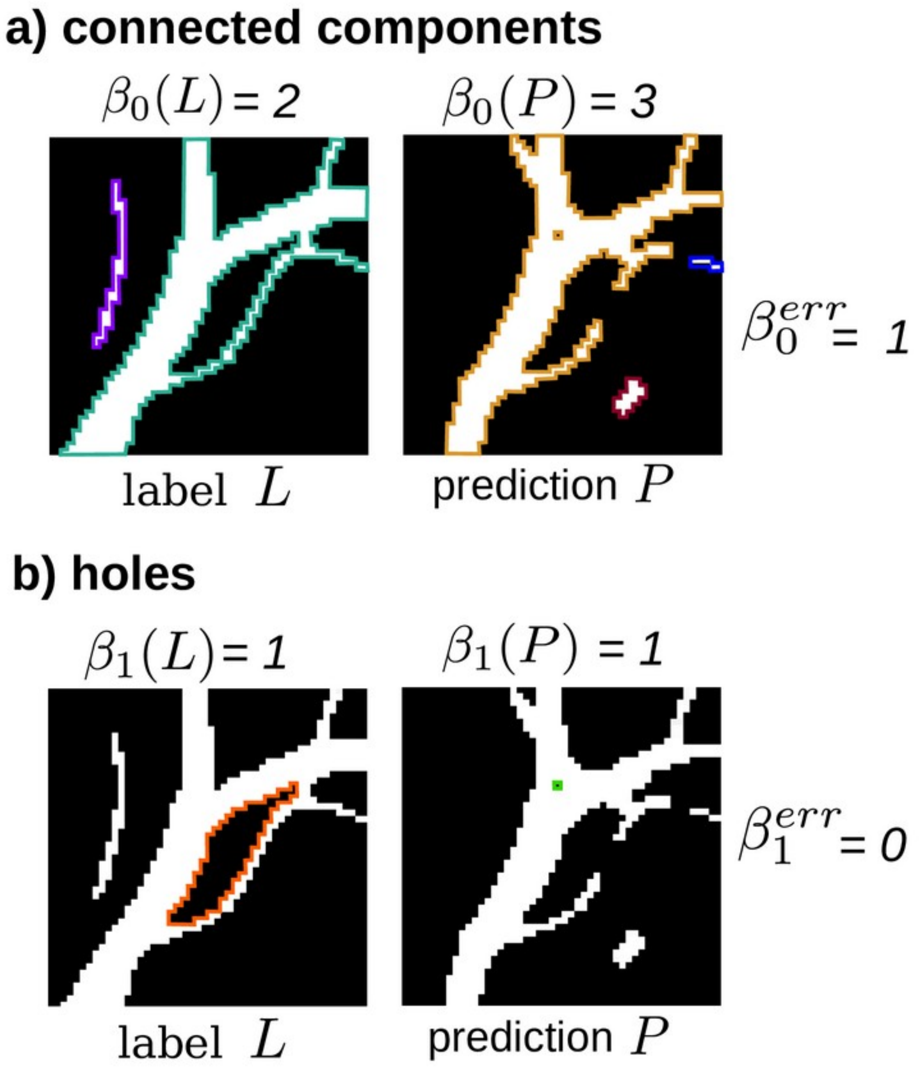
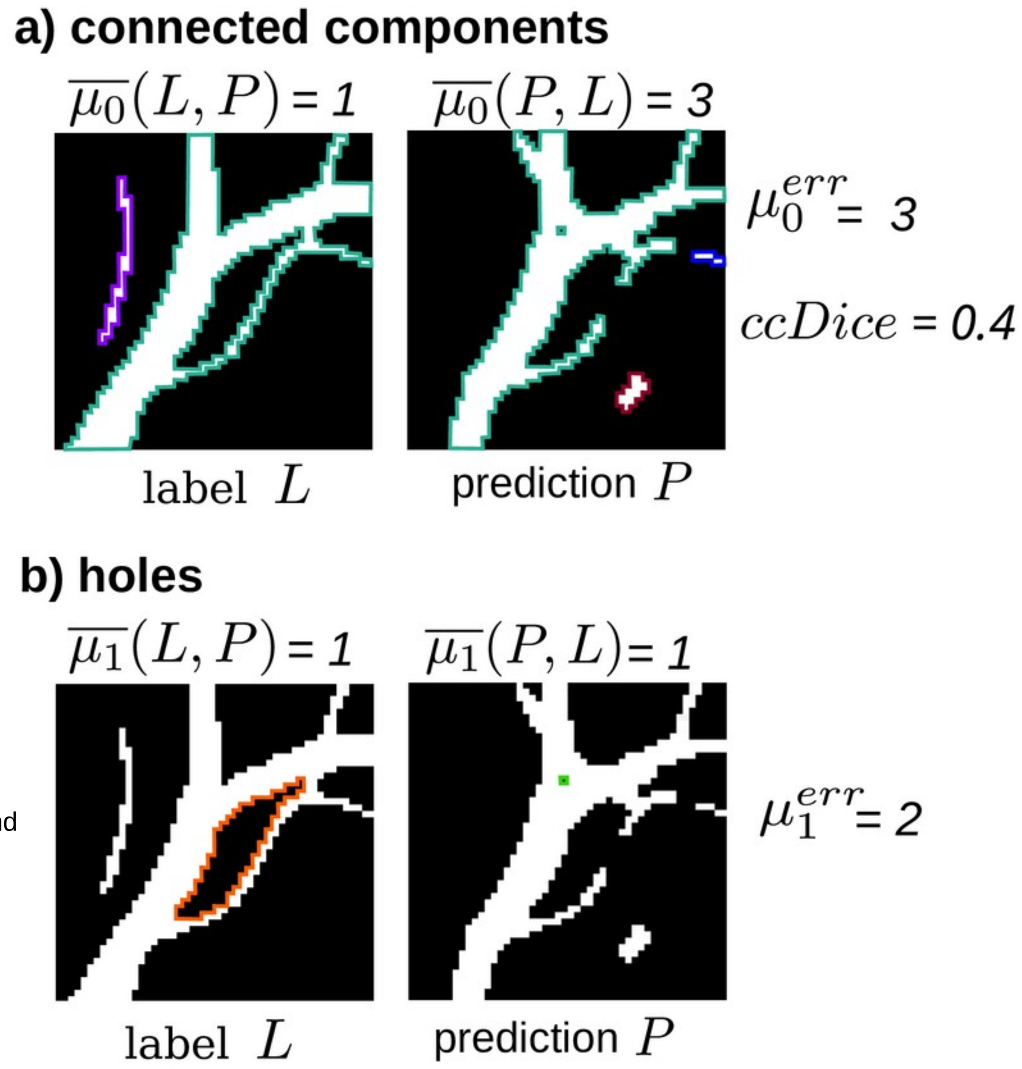


Figure 6. Illustration of the Betti error calculation.

Introduction – Algebraic topology quality metrics

- The **Betti matching error**[3] μ^{err} and the **ccDice**[4] are matching CC and holes.



[3]Stucki, Nico, et al. "Topologically faithful image segmentation via induced matching of persistence barcodes." International Conference on Machine Learning, 2023.

[4]Rougé, Pierre, Odyssée Merveille, and Nicolas Passat. "ccDice: A topology-aware Dice score based on connected components." International Workshop on Topology-and Graph-Informed Imaging Informatics, MICCAI 2024.

Figure 7. Illustration of the Betti matching and ccDice calculation.

Challenge

- Which metric / loss shall I use? How to be sure than the metric reflects my expectations?
  → Quality is **subjective**, making it difficult to assess the **strengths and weaknesses** of each metric.

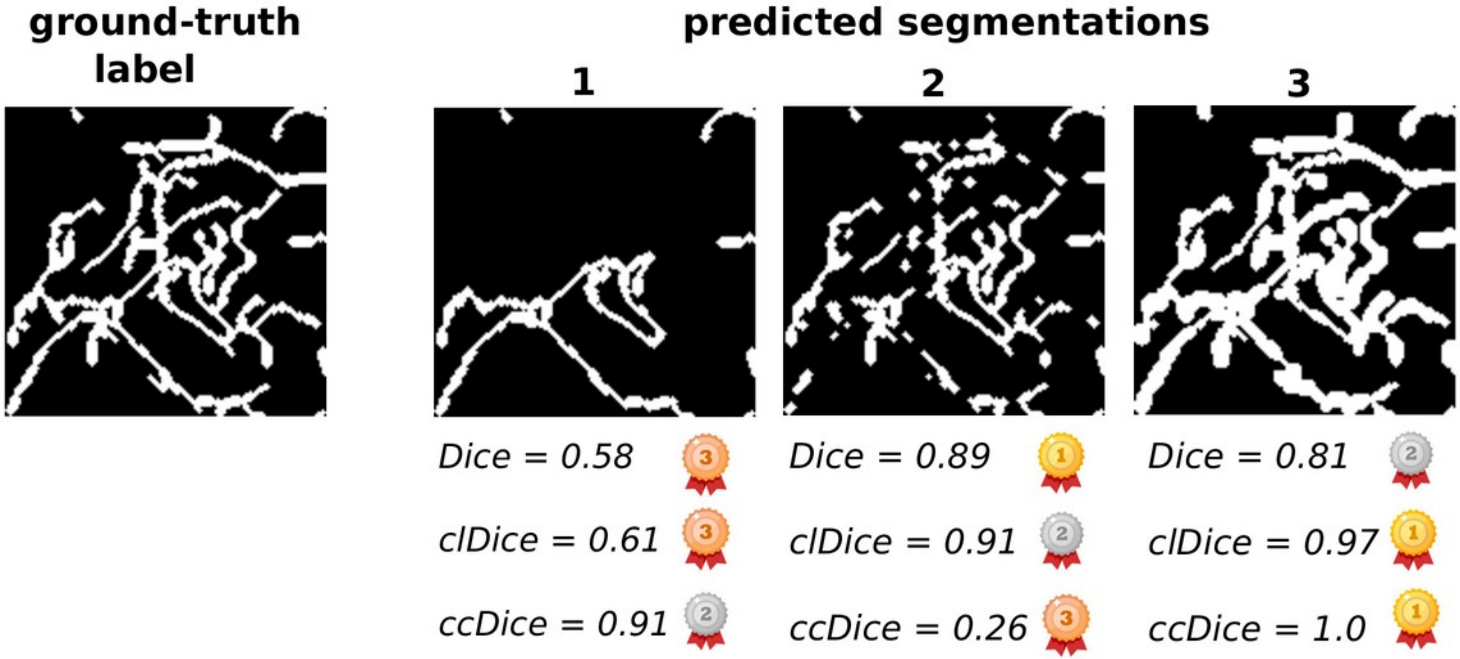


Figure 8. Which is the best segmentation?

State-of-the-art

How the evaluate quality metrics?

1) Use handcrafted examples to show the metric behavior in specific cases[2,4]

2) Ask experts to identify pitfalls and make recommendations (*Metrics Reloaded project*)[5,6]

3) Evaluate metrics based on their correlation to subjective visual scores attributed by experts[7]

[5]Maier-Hein, Lena, et al. "Metrics reloaded: recommendations for image analysis validation." Nature methods, 2024
[6]Reinke, Annika, et al. "Understanding metric-related pitfalls in image analysis validation." Nature methods, 2024
[7]Aydin, Orhun Utku, et al. "An evaluation of performance measures for arterial brain vessel segmentation." BMC medical imaging, 2021

State-of-the-art

How the evaluate quality metrics?

1) Use handcrafted examples to show the metric behavior in specific cases[2,4]

2) Ask experts to identify pitfalls and make recommendations (*Metrics Reloaded project*)[5,6]

3) Evaluate metrics based on their correlation to subjective visual scores attributed by experts[7]

Limitations

- No study focusing on topology-preserving metrics so far
- Relying on experts to grade images or identify pitfalls : time-consuming, hard to apply to new metrics
- Limited to the specific contexts, such as particular application or dataset.

State-of-the-art

How the evaluate quality metrics?

1) Use handcrafted examples to show the metric behavior in specific cases[2,4]

2) Ask experts to identify pitfalls and make recommendations (*Metrics Reloaded project*)[5,6]

3) Evaluate metrics based on their correlation to subjective visual scores attributed by experts[7]

Limitations

- ~~No study focusing on topology-preserving metrics so far~~
- ~~Relying on experts to grade images or identify pitfalls : time-consuming, hard to apply to new metrics~~
- ~~Limited to the specific contexts, such as particular application or dataset.~~

→ We propose a new approach to benchmark metrics without the need for expert knowledge!

Proposed approach

- Depending on the application, different segmentation errors have different importance or “**weight**”.

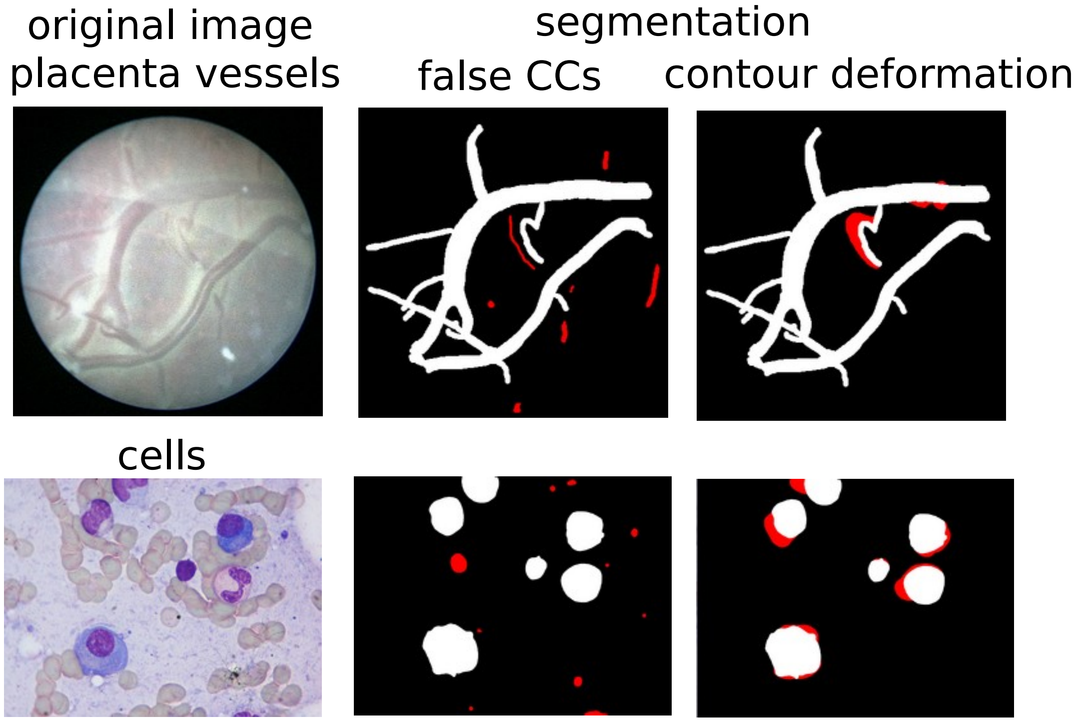


Figure 9. Illustration of different types of segmentation errors. Images from the PSVFM[8] and Neurips 2022 dataset[9].

[8]Bano, Sophia, et al. "Deep placental vessel segmentation for fetoscopic mosaicking." MICCAI, 2020.
[9]Ma, Jun, et al. "The multimodality cell segmentation challenge: toward universal solutions." Nature methods, 2024

Proposed approach

→ **Break down** error types into **easily interpretable categories** (e.g. false component, deformation…)

→ Estimate the “**weight**” that a given metric attributes to each type of errors.

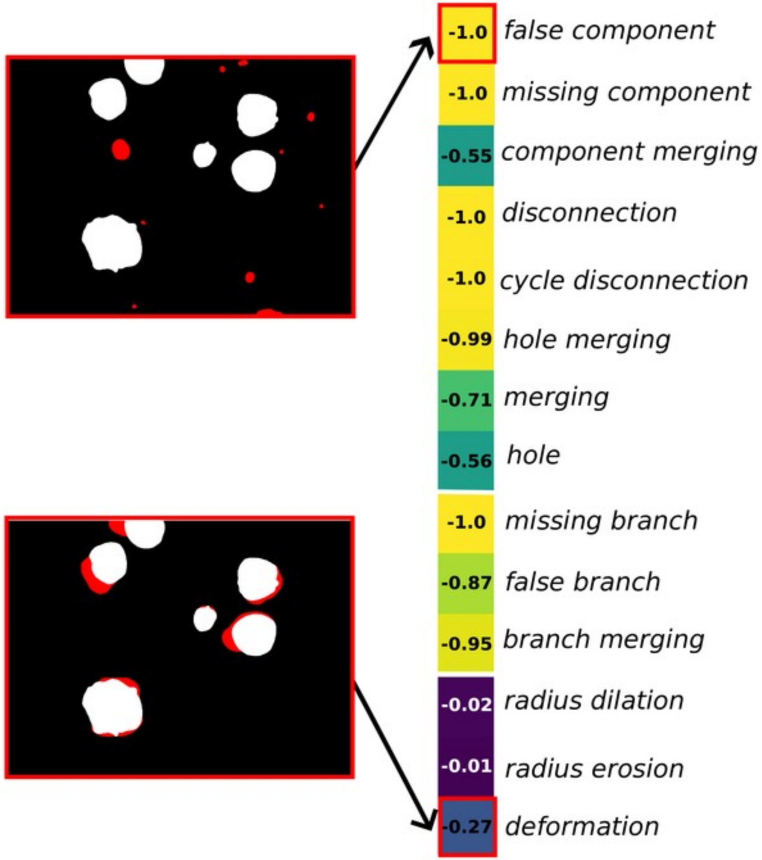


Figure 10. Estimated weights for each type of segmentation errors.

Proposed approach

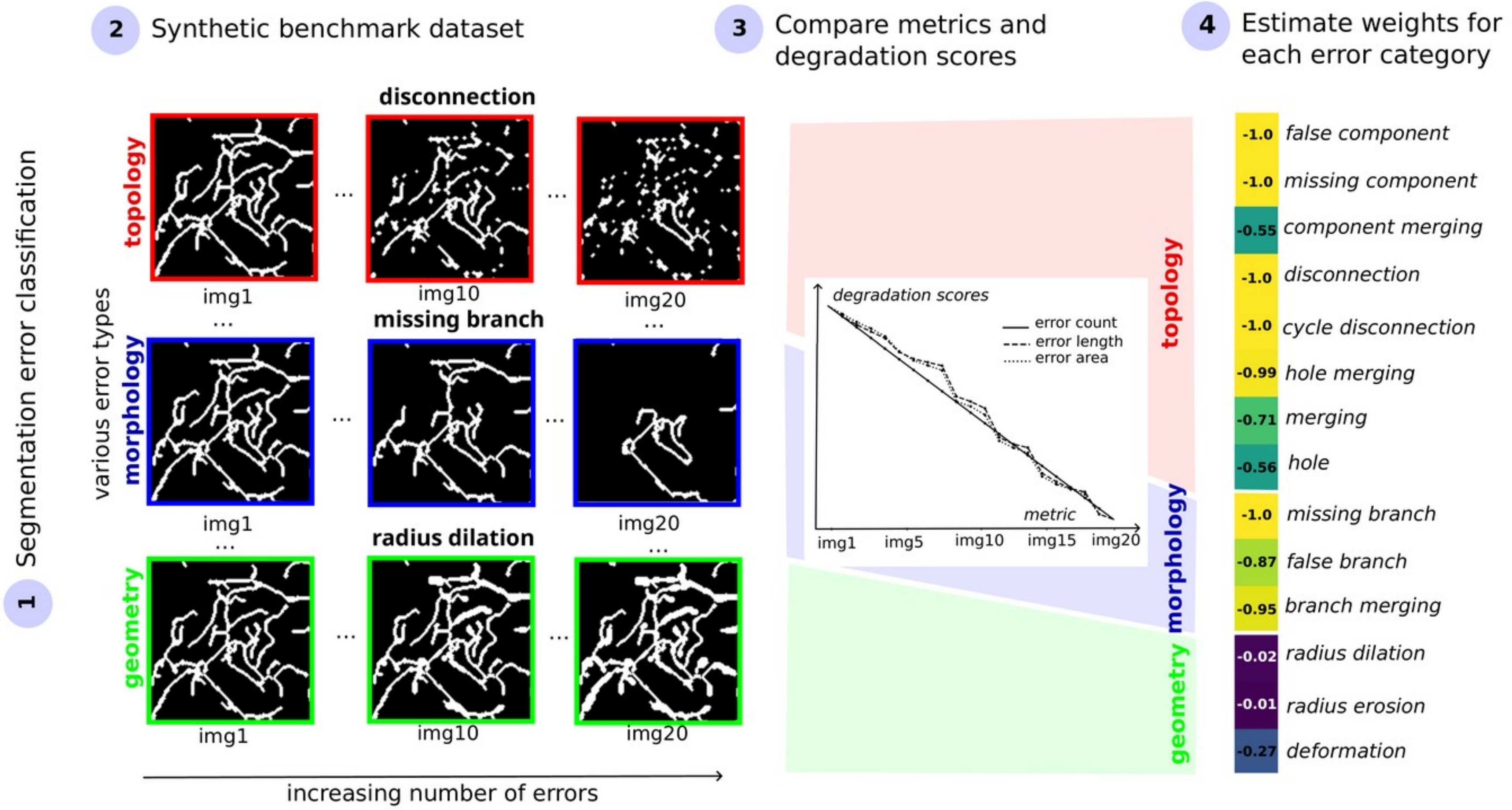


Figure 11. Overview of the proposed method.

Method – Error classification

- We introduce a new distinction between "**topology**", "**morphology**" and "**geometry**".

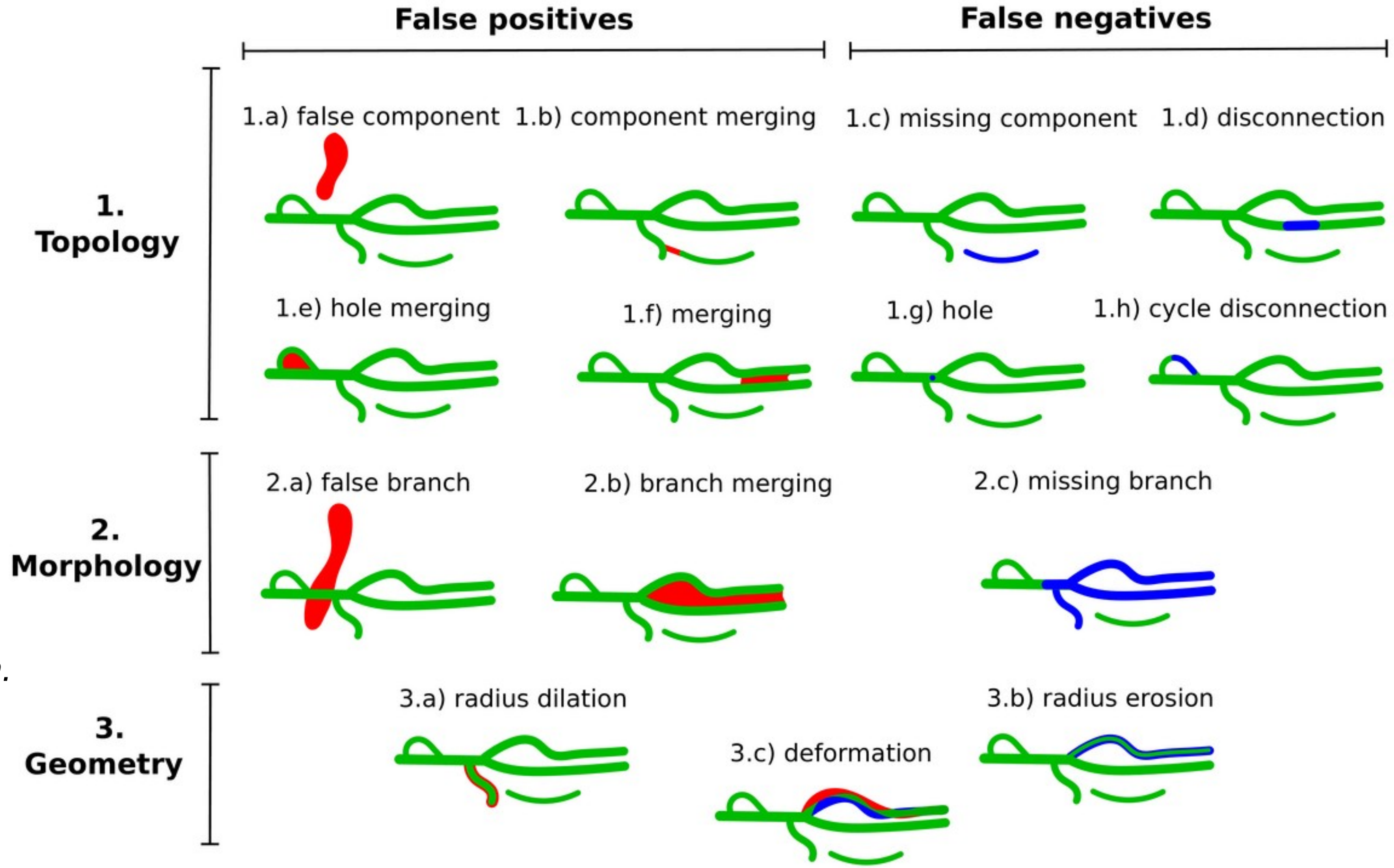
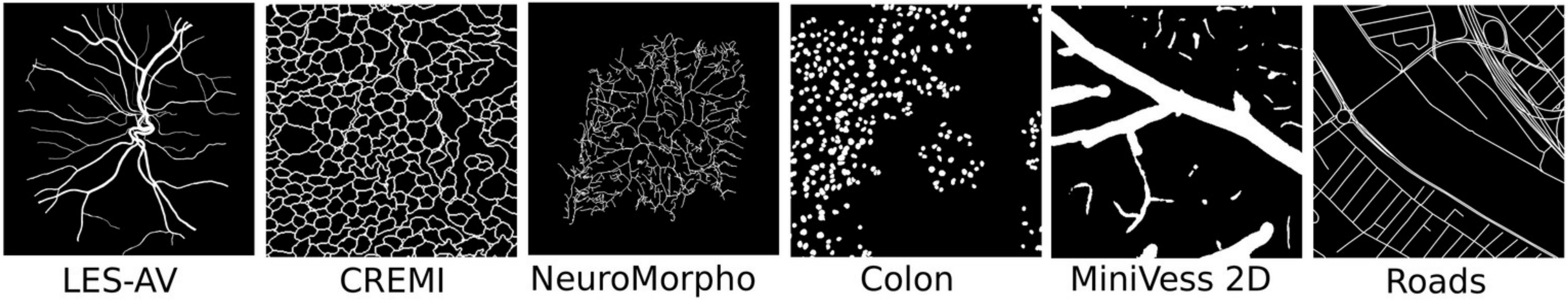


Figure 12. The proposed Segmentation error classification.

Method – Synthetic benchmark dataset

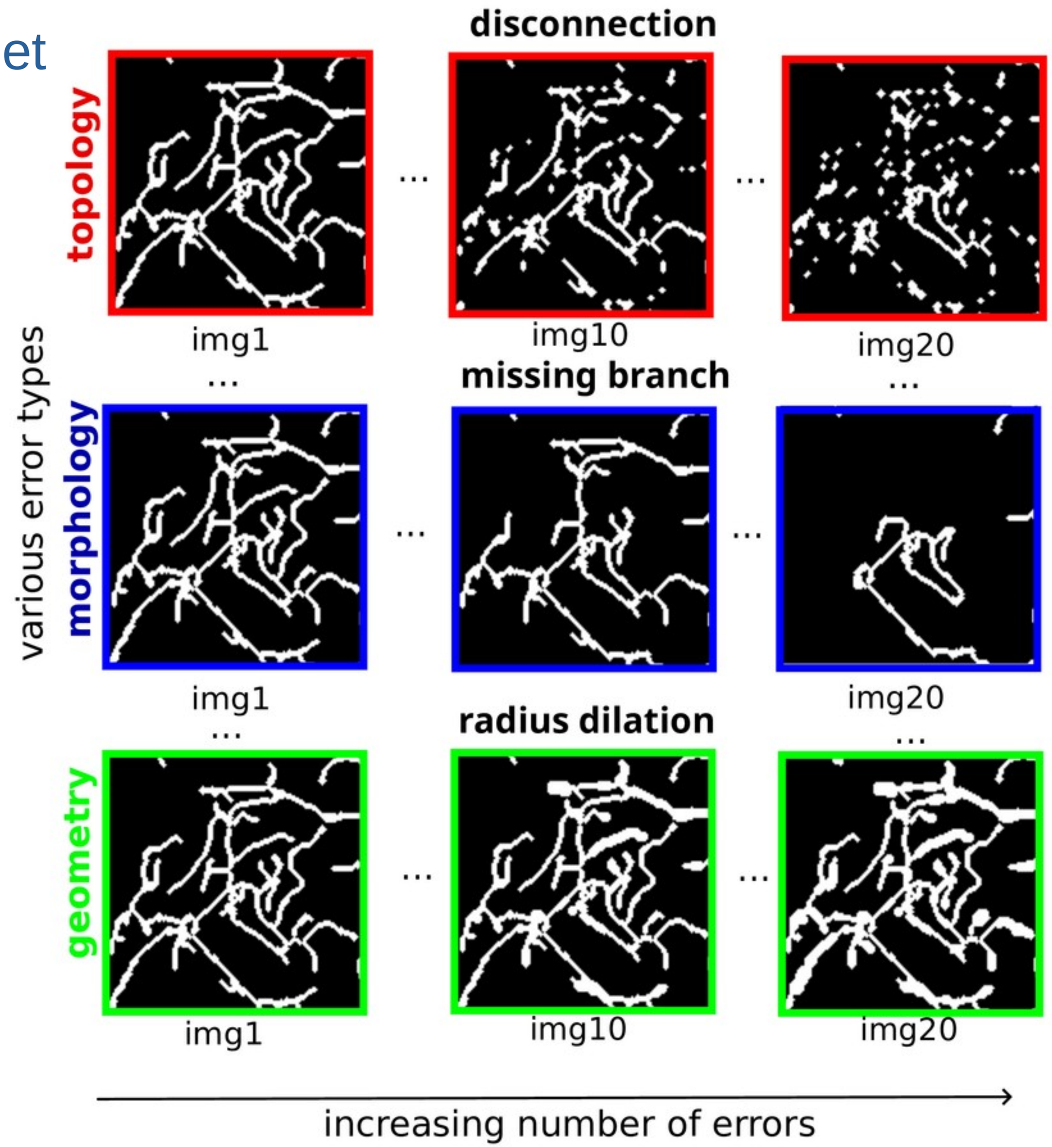
- We selected 3 labels from 6 public datasets.



[10]Cremi. miccai challenge on circuit reconstruction from electron microscopy images https://cremi.org/ (2016), accessed: 2024-02-12
[11]Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature methods, 2012
[12]Tecuatl, C., Ljungquist, B., Ascoli, G.A.: Accelerating the continuous community sharing of digital neuromorphology data. FASEB BioAdvances, 2024
[13]Mnih, V.: Machine Learning for Aerial Image Labeling. Ph.D. thesis, University of Toronto, 2013
[14]Odstrcilik, J., Kolar, R., Budai, A., Hornegger, J., Jan, J., Gazarek, J., Kubena, T., Cernosek, P., Svoboda, O., Angelopoulou, E.: Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Processing, 2013
[15]Poon, C., Teikari, P., Rachmadi, M.F., Skibbe, H., Hynynen, K.: A dataset of rodent cerebrovasculature from in vivo multiphoton fluorescence microscopy imaging. Scientific Data, 2023

Method – Synthetic benchmark dataset

- We **cumulatively add errors** from each category to mimic predicted segmentations.



Method – Degradation scores

- We design **degradation scores** to quantify the degradation of synthetic images. We consider three **error properties** : the error *“area”*, the error *“length”* and the error *“count”*.

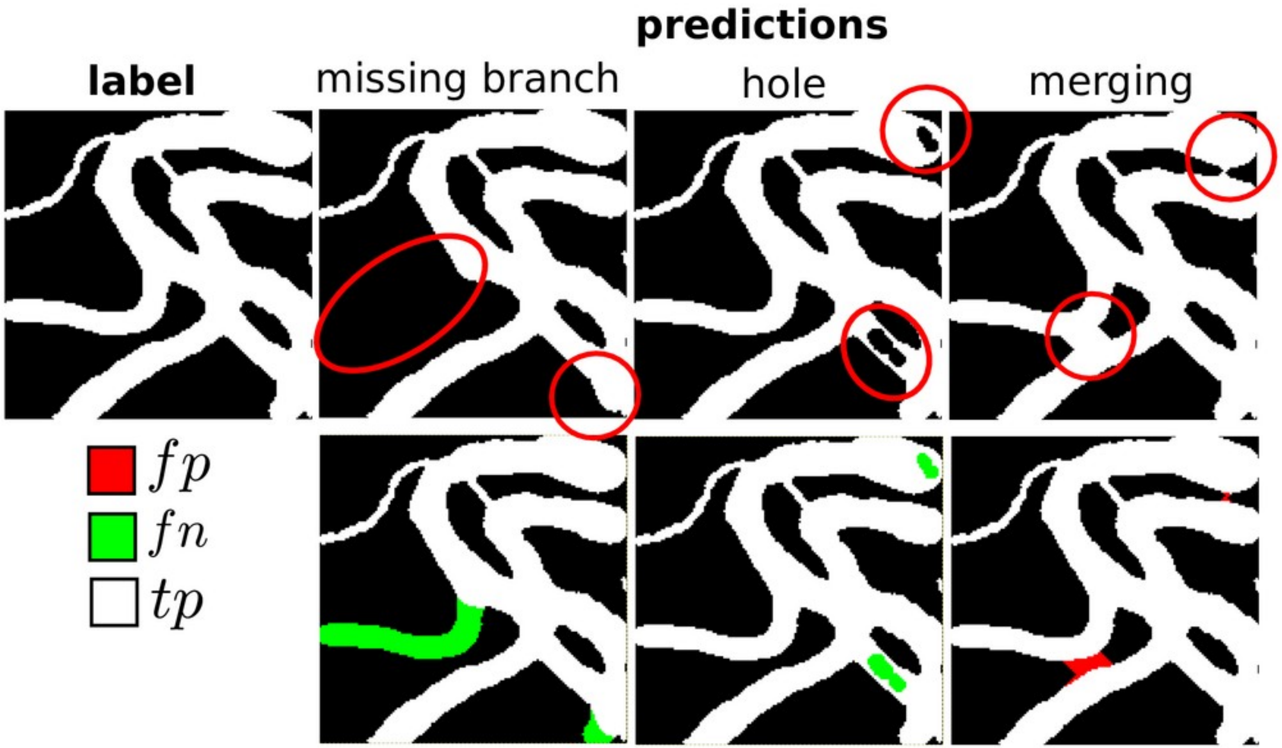


Figure 13. Calculation of the degradation scores for the property “area”.

Method – Degradation scores

- We design **degradation scores** to quantify the degradation of synthetic images. We consider three **error properties** : the error *“area”*, the error *“length”* and the error *“count”*.

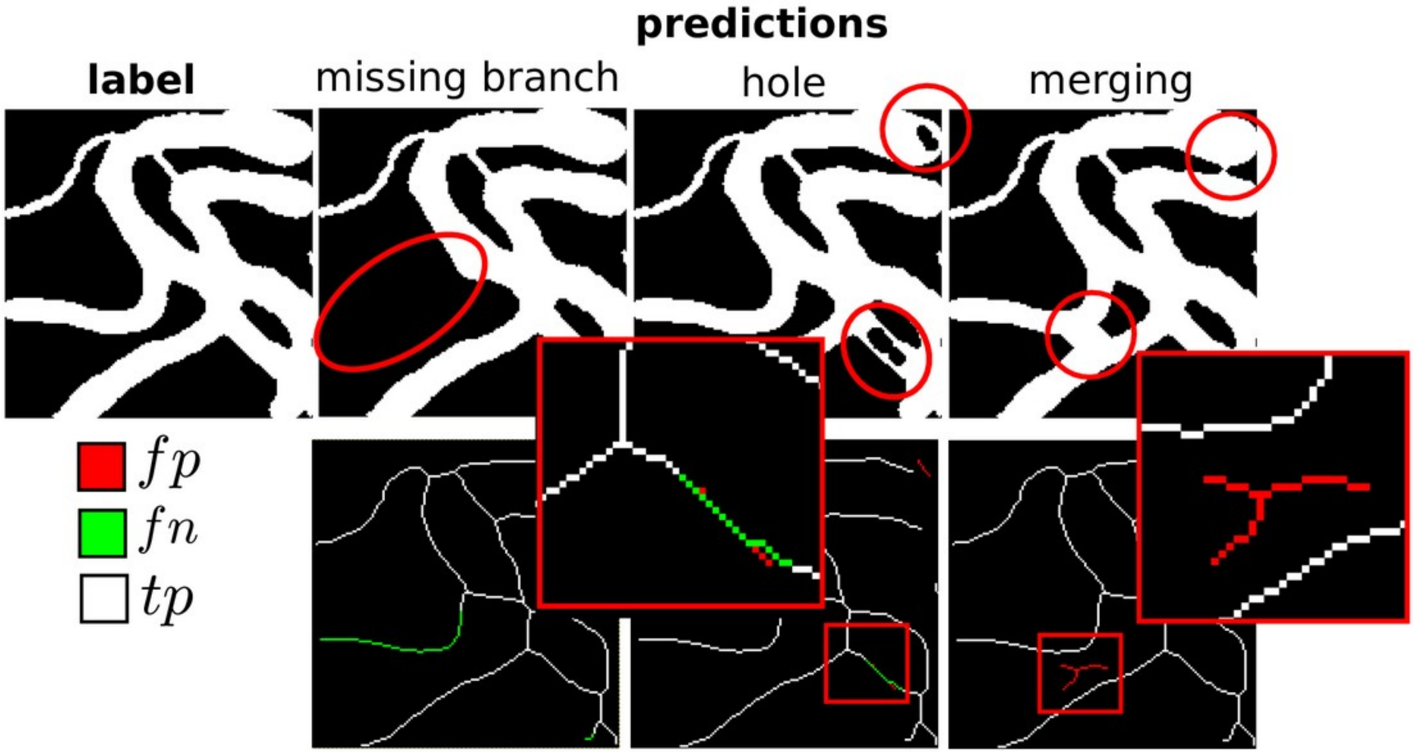
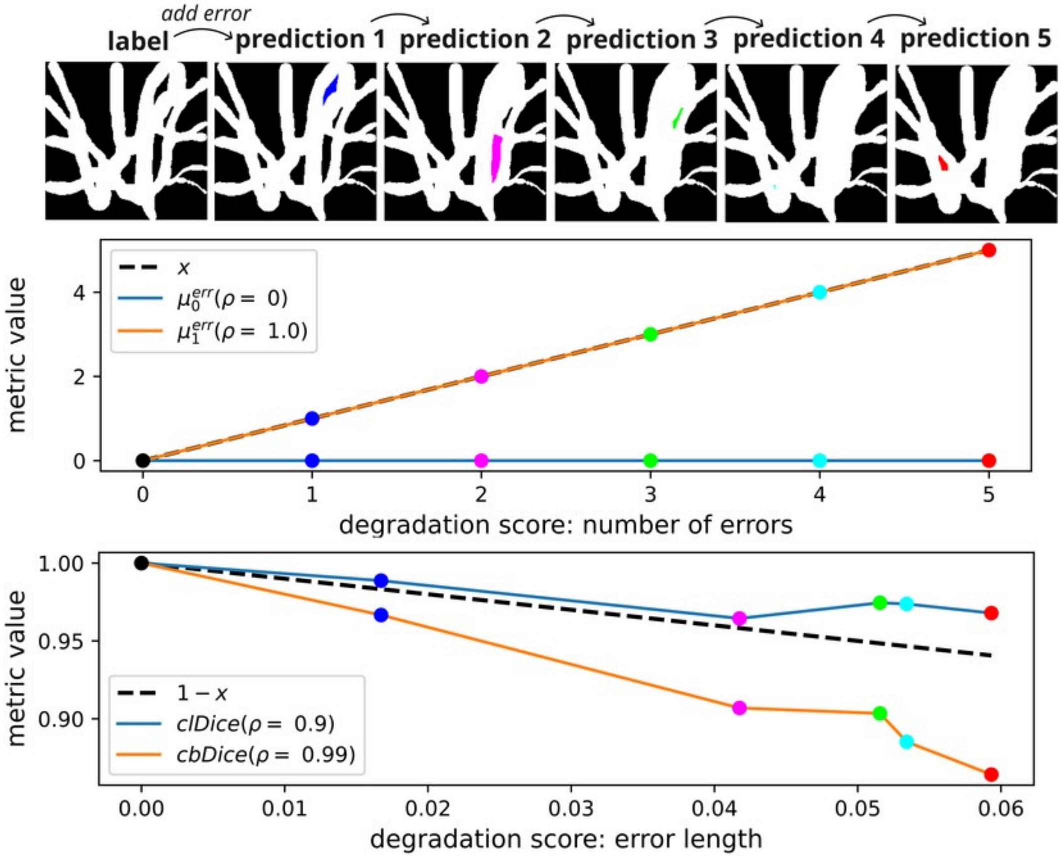


Figure 14. Calculation of the degradation scores for the property “length”.

Method – Correlation and weights estimation



- **Correlation** (absolute Pearson coefficient)

= Is the metric sensitive to one type of error (ρ = 1) or not (ρ = 0)?

- **Weight** (slope of fitted linear function)

= How strongly does the metric penalize a a given type of error?

Results – Correlation analysis

(ρ = 0) : Does not account for this type of error
(ρ > 0.9) : Good estimator of the type of error
(ρ < 0.9) : Influenced by the error (limitation?)

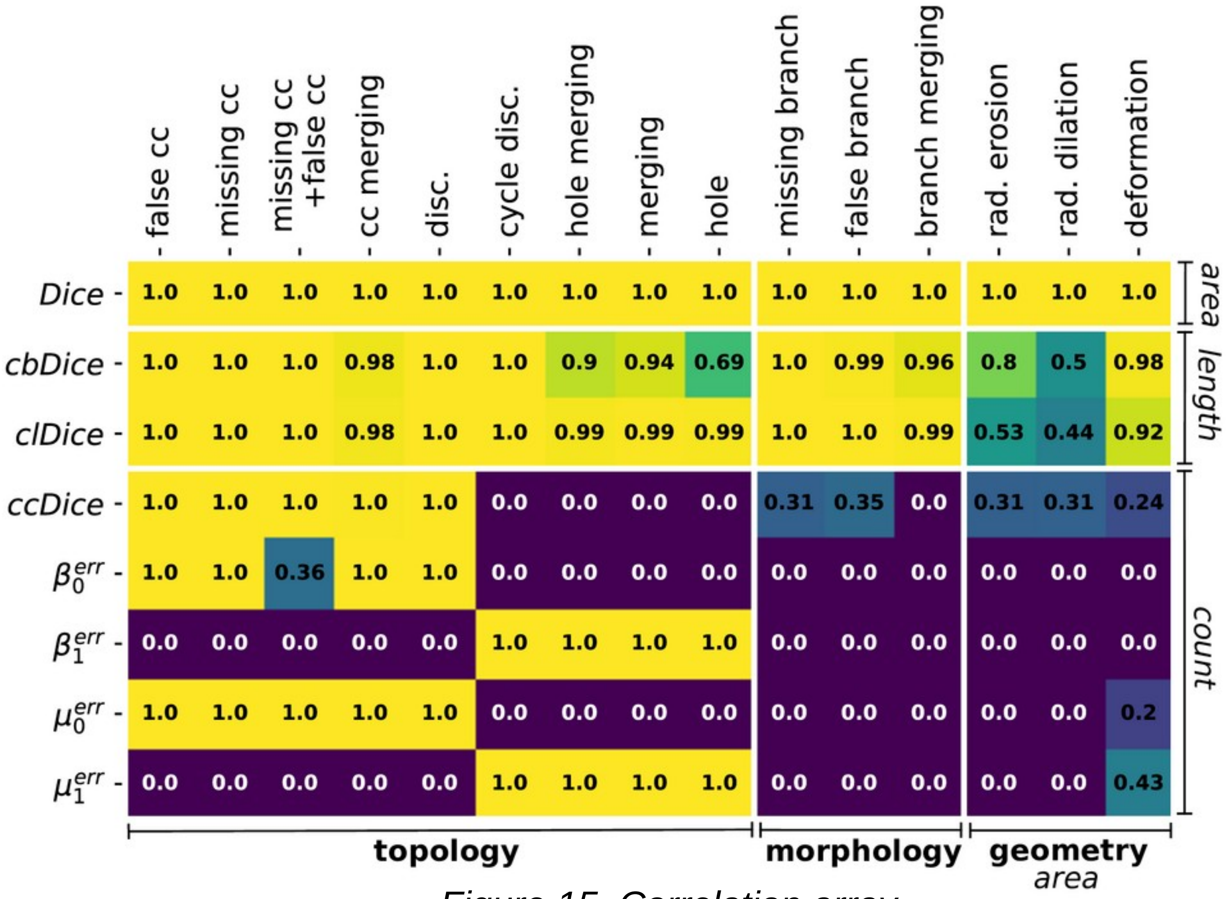


Figure 15. Correlation array.

Results – Correlation analysis

(ρ = 0) : Does not account for this type of error
(ρ > 0.9) : Good estimator of the type of error
(ρ < 0.9) : Influenced by the error (limitation?)

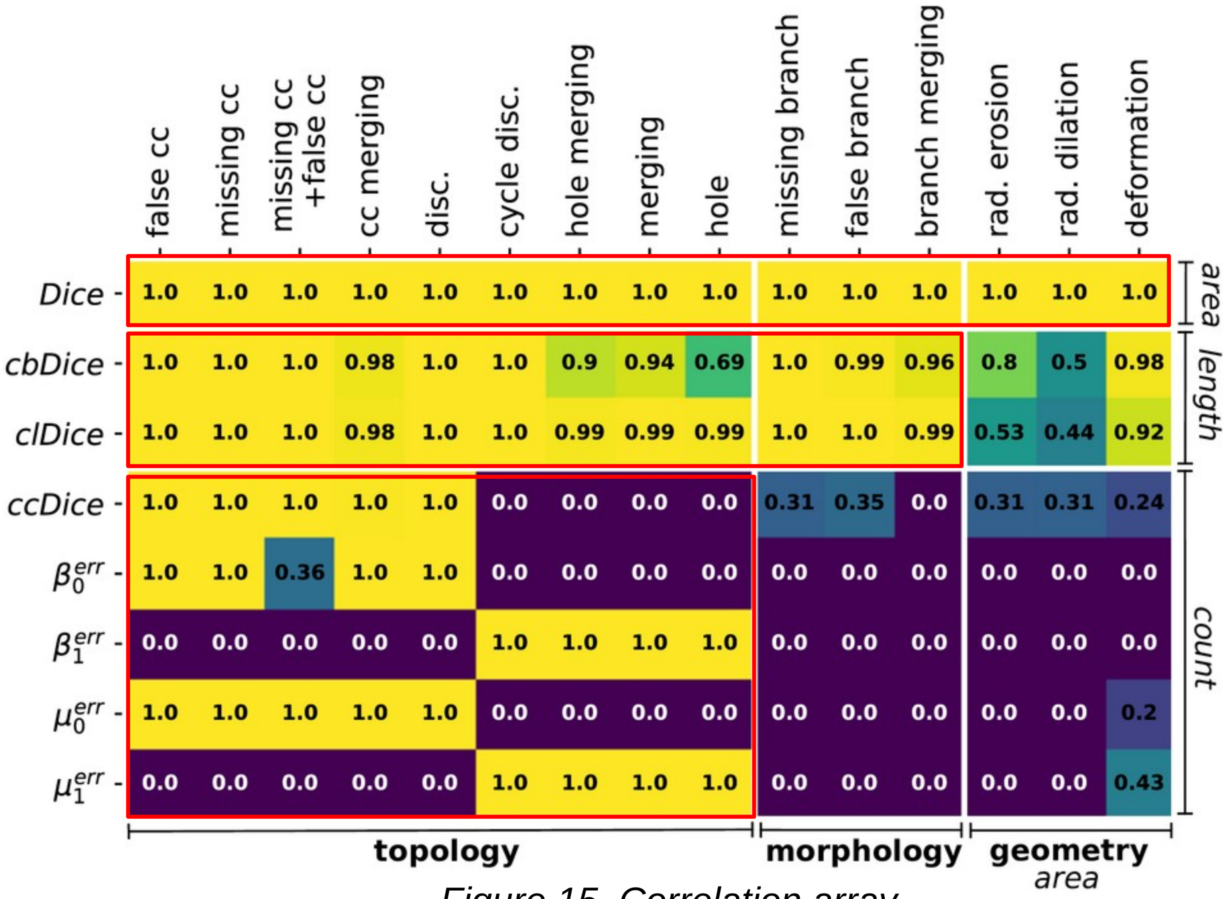


Figure 15. Correlation array.

Results – Correlation analysis

- The clDice is affected by **radius change and deformations**.

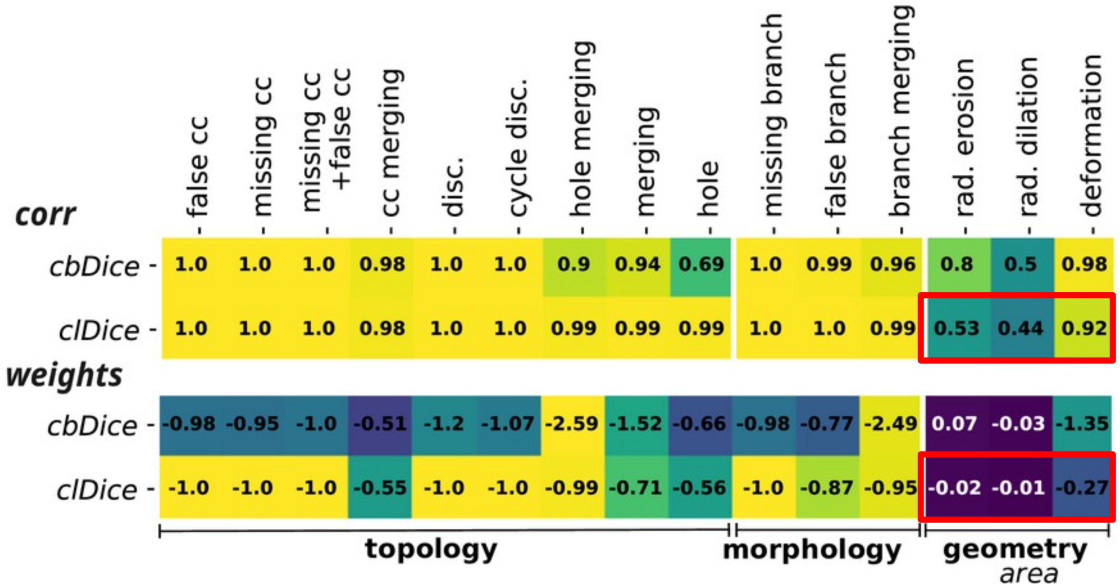


Figure 16. Correlations and weights for the clDice and cbDice.

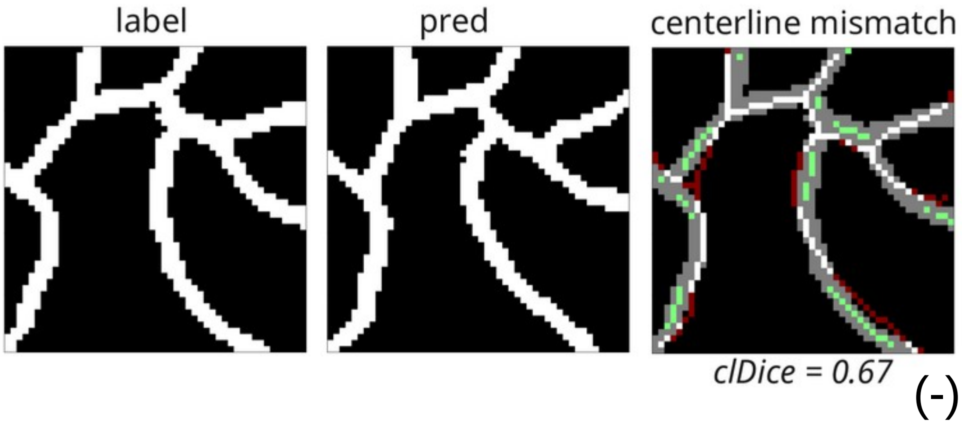


Figure 17. Illustration of the clDice behavior.

Results – Correlation analysis

- The clDice may overlook **holes and merging.**

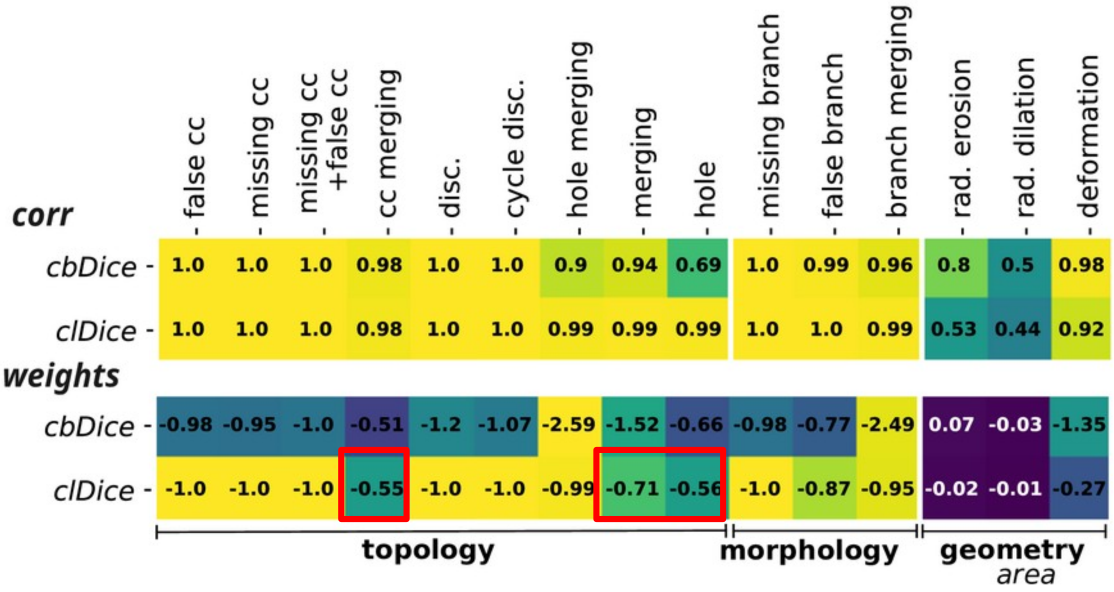


Figure 18. Correlations and weights for the clDice and cbDice.

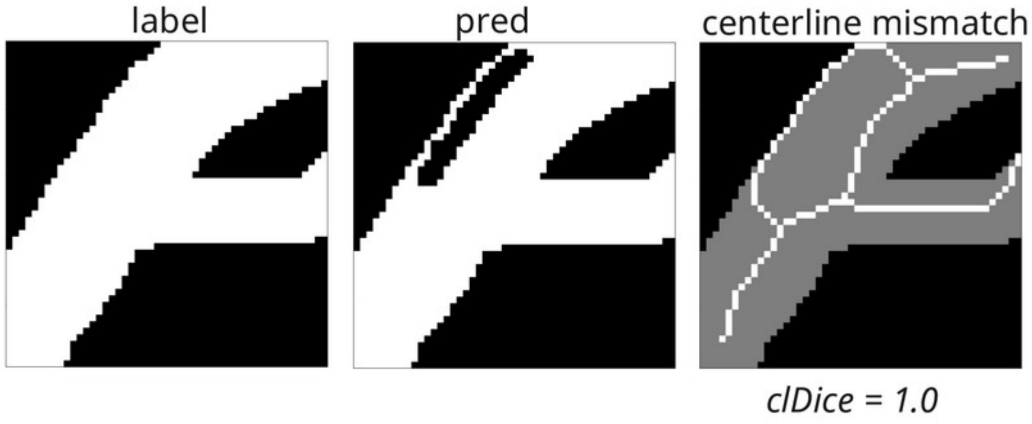


Figure 19. Illustration of the clDice behavior.

(++)

Results – Correlation analysis

- The ccDice and Betti metrics **ignores morphological errors** (e.g. missing branches)

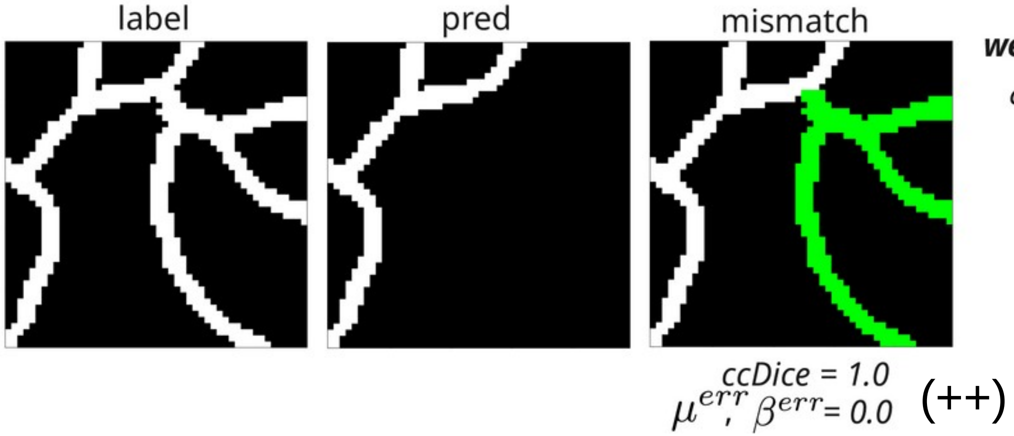


Figure 20. Illustration of the metric's behavior.

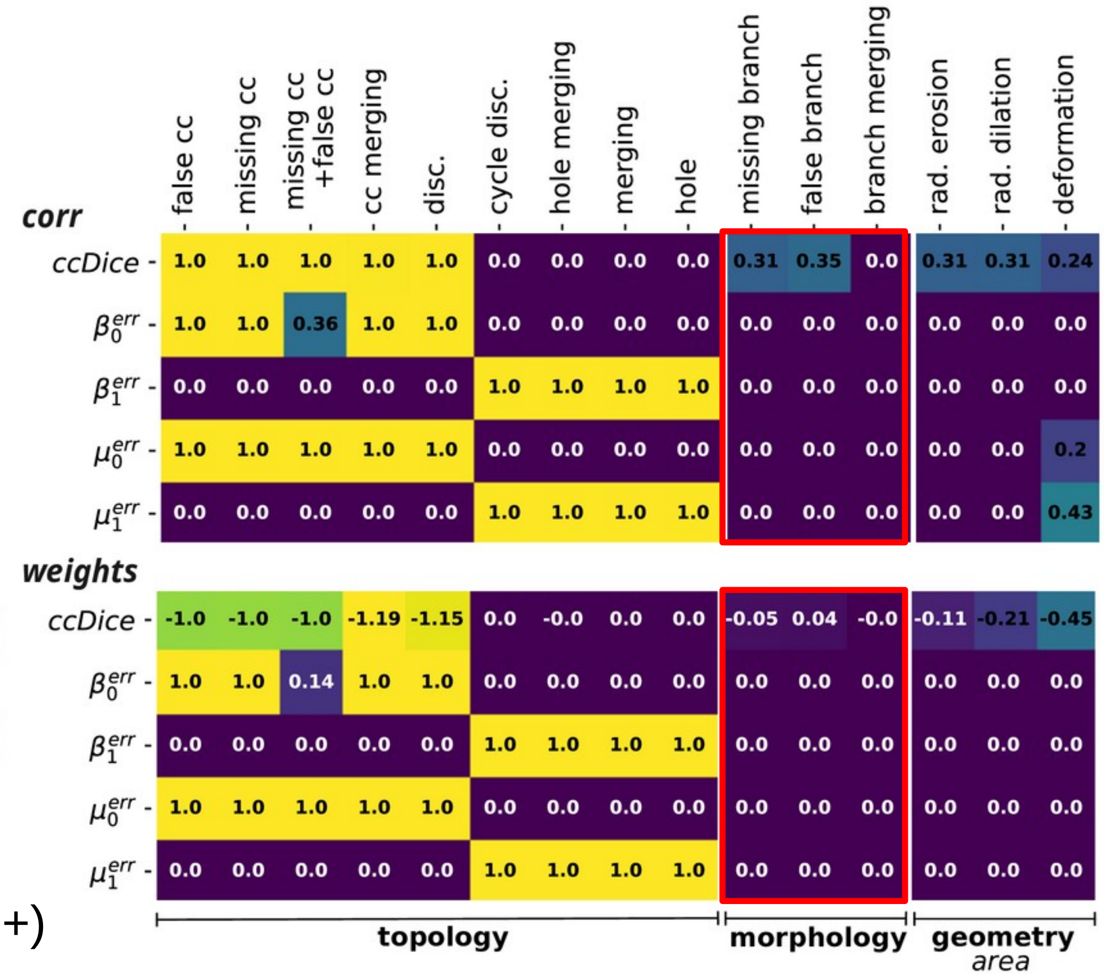


Figure 21. Correlations and weights for the ccDice and Betti metrics.

Conclusion

Contributions

- First classification of segmentation errors for tubular structures
- Method to generate synthetic segmentation with a given type of error
- New approach to visualize and interpret the metrics behavior

Conclusion

Contributions

- First classification of segmentation errors for tubular structures
- Method to generate synthetic segmentation with a given type of error
- New approach to visualize and interpret the metrics behavior

Advantages +

- No need for expert knowledge on the metrics
- Automated (easy to apply to new metrics)
- Covers a large range of applications and contexts (can find unexpected pitfalls!)

Conclusion

Contributions

- First classification of segmentation errors for tubular structures
- Method to generate synthetic segmentation with a given type of error
- New approach to visualize and interpret the metrics behavior

Advantages +

- No need for expert knowledge on the metrics
- Automated (easy to apply to new metrics)
- Covers a large range of applications and contexts (can find unexpected pitfalls!)

Limitations -

- Degradation scores may not reflect the desired metric behavior.
- Necessary to consider more error properties (boundary, center-of-mass)