Multi-modal segmentation for paramagnetic rim lesion detection in multiple sclerosis

Maxence Wynen[a,b], Pedro M. Gordaliza[c], Anna Stölting[b], Pietro Maggi[b,c], Meritxell Bach Cuadra[d,e], and Benoit Macq[a]

[a]ICTeam, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[b]Louvain Neuroinflammation Imaging Lab (NIL), Université Catholique de Louvain, Brussels, Belgium
[c]Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Brussels, Belgium
[d]CIBM Center for Biomedical Imaging, Switzerland
[e]Department of Radiology, Lausanne University Hospital, Lausanne, Switzerland

1. INTRODUCTION

Multiple Sclerosis (MS) is a prevalent chronic autoimmune disorder affecting millions worldwide. It is characterized by focal lesions in the central nervous system, notably in the brain white and gray matter, along with the spinal cord. Magnetic Resonance Imaging (MRI) has emerged as a crucial tool for diagnosing[1] and predicting disease progression[2] in MS patients. The current MRI criteria, however, still lack specificity.[3] To overcome these limitations, novel biomarkers have been explored through recent advancements in MR technology and clinical research studies.[4] In particular, the investigation of Paramagnetic Rim Lesions (PRL) holds promise to assist both in diagnosing MS and in monitoring disease progression over time.[5] PRL (Figure 1) are chronic-active lesions characterized by a rim of iron-laden (pro-)inflammatory cells around their core in susceptibility-weighted imaging (SWI). Importantly, white matter (WM) lesions have the potential to merge, resulting in the formation of "confluent lesions" (Figure 1 II). This merging phenomenon makes it hard to accurately count lesions for prognostic purposes. In this context, the visualization of rims through SWI facilitates the differentiation of confluent lesions, enabling a more precise determination of lesion count and the associated clinical implications. These attributes set PRL apart as a potential unique marker for MS diagnosis/prognosis, differentiating them from other WM lesions.

The manual assessment of PRL is laborious, susceptible to substantial annotation bias[6] and prone to high inter-rater variability.[7] Hence, the development of an automatic assessment method holds vital importance. Given the novelty of the biomarker and the task complexity, only three methods[8,9,7] aiming to detect PRL have been published to date. These methods follow a two-stage process: the first involves either manual or automatic lesion delineation based on a precomputed lesion segmentation, referred to as lesion-splitting, followed by the classification and labelling of lesions employing common machine learning approaches. However, the lesion-splitting step encounters two main challenges:[10] the potential for non-PRL imaging manifestations to mimic PRLs in common clinical sequences (Figure 1), and the presence of confluent lesions making their isolation a challenging task. This is particularly relevant, as PRL often constitute confluent lesions due to their smouldering and expanding nature through the WM tissue,[7] thereby intensifying the intricacies of detection.

To address these challenges, our work introduces the first approach to direct PRL segmentation as a tool to highlight potential regions containing PRL, thus removing the need of lesion splitting. By adopting this novel strategy, we aim to overcome the complexities posed by confluent lesions and demonstrate the potential utility of a straightforward PRL segmentation as well as their challenges. While our study's results are preliminary, they provide valuable insights and lay the foundation for future research in this unexplored area. Emphasizing the significance of developing an automatic PRL segmentation method, our work holds the potential to significantly enhance MS diagnosis and monitoring. The insights gained could offer clinicians new information that directly impacts patient care and treatment development.

Further author information: (Send correspondence to maxence.wynen@uclouvain.be)

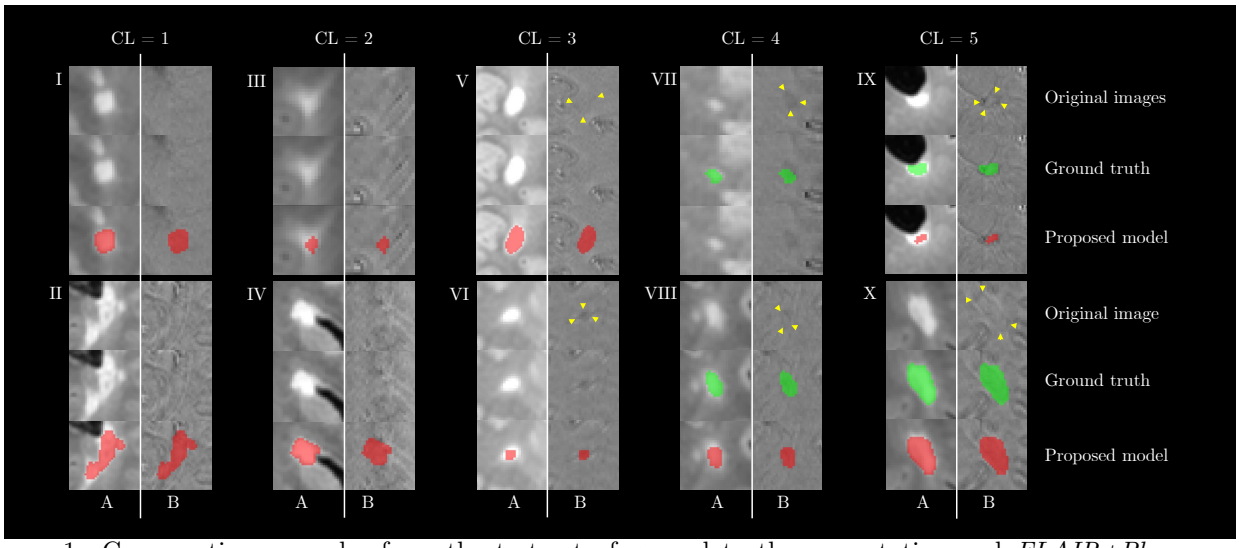

Figure 1: Comparative examples from the test set of ground truth segmentation and *FLAIR+Phase* model predictions for Paramagnetic Rim Lesions (PRL). FLAIR (A) and T2*-EPI Phase (B) images are shown. Visual Confidence Levels (CL) range from 1 (non-PRL) to 5 (PRL). Examples I-VI depict False Positives, VII a False Negative, and VIII-X True Positives. Yellow arrows indicate the presence of a rim.

2. MATERIALS

This study enrolled 47 subjects (26 female, 21 male), aged 40±12.7. The distribution of MS clinical subtypes was as follows: *Relapsing-Remitting MS* (RRMS) - 32, *Secondary Progressive MS* (SPMS) - 11, *Primary Progressive MS* (PPMS) - 3, and *Radiologically Isolated Syndrome* (RIS) - 1. The inclusion criteria for subject selection mandated that all participants had at least one PRL visible on their MRI scans.

All subjects underwent MRI scanning at Saint-Luc University Hospital in Brussels, Belgium using a 3T Signa Premier General Electrics MRI scanner. The MRI protocol included the acquisition of three key 3D sequences: *Fluid-Attenuated Inversion Recovery* (FLAIR, $1mm^3$), *Magnetization-Prepared Rapid Acquisition Gradient Echo* (MPRAGE, $1mm^3$), and *T2*-weighted Echo Planar Imaging* (T2*-EPI, $0.667mm^3$) magnitude and unwrapped phase which will hereafter be referred as Magnitude and Phase respectively.

To establish ground truth labels for the PRL, a consensus annotation was obtained from two experts, P.M. and A.S., using all images at their disposal. The experts provided visual confidence levels for each lesion using a Likert scale ranging from 1 to 5. A confidence level of 1 indicated absolute certainty that the lesion was not a PRL, while a confidence level of 5 indicated absolute certainty that the lesion was a PRL (see Figure 1). Only lesions with a confidence level of 4 or higher were considered as PRL for the study. Moreover, only deep WM lesions were included in the dataset, while cortical and infratentorial lesions, as well as glioses, vascular damage, and tumefactive lesions, were excluded.

3. METHODS

3.1 Preprocessing

All images underwent masking and cropping using the brain mask provided by *FSL-Brain Extraction Tool*.[11]

As highlighted in the literature on CNN-based lesion segmentation studies,[12] there is potential benefit in utilizing an attention mechanism by supplying models exclusively with the region of interest. In our case, this attention mechanism is directed towards the healthy or damaged WM parenchyma. Therefore, we conduct experiments employing two distinct preprocessing strategies:

1. Applying a pre-computed WM lesion mask, generated with *Samseg*[13] and subsequently dilated to encompass lesion borders. This mechanism emphasizes the attention to areas surrounding the lesions.

2. Applying a WM mask computed by *FreeSurfer*,[14] lesion-filled with the dilated lesion segmentation from *Samseg*.[13] This aims to drive the network to exclusively focus on WM while retaining pertinent lesion information.

3.2 Model Architecture and Training

For our analysis, we employed a 3D-UNet architecture[15] with a binary output, distinguishing between PRL segmentation and background. To ensure the appropriateness of this baseline model, we conducted a preliminary benchmark on our dataset using the nnUNet[16] framework. The consistent results align with existing literature, confirming that the 3D-UNet architecture is a fitting initial model. To establish a standardized training approach, we adopted the WM lesion segmentation ad-hoc baseline from the MS Shift Challenge 2.0.[17] Moreover, since this study marks a pioneering effort in PRL segmentation, it is of vital importance to assess the capabilities of the various available MRI sequences, as well as their combined information, to instruct the model in distinguishing the biomarker of interest from the rest of the parenchyma.

To this aim, the model was trained with one modality as input for every available modality. In addition, to leverage the complementary information from different modalities, we strategically concatenated select modalities at the network input. Specifically, we combined relevant modalities after discussion with experts, including 1) *FLAIR+Phase*, 2) *FLAIR+MPRAGE*, and 3) *FLAIR+Phase+MPRAGE*. Regarding training details, all model variations were trained using a learning rate of 1e-4 for 300 epochs, employing a cosine annealing learning rate scheduler to optimize convergence. The training loss consisted of a weighted combination of dice loss[18] and focal loss.[19] For evaluation, we randomly divided the 47 subjects into three sets: 28 subjects for training (282 PRL), 9 subjects for validation (46 PRL), and 10 subjects for testing (56 PRL).

3.3 Evaluation Metrics

Given the sparse nature of PRL, we have chosen to evaluate our study through the lens of instance segmentation. While our work involves semantic segmentation, using metrics aligned with that approach could potentially yield misleading interpretations. These metrics often lack precision and are not well-suited for our medical context, a fact underscored by the existing literature.[20] To overcome this challenge, we have embraced the recommendations of the Metrics Reloaded framework for assessing sparse lesions. This methodology involves isolating the regions detected by our model through the application of the full connected components algorithm. Subsequently, we compare these isolated regions with our reference using boundary Intersection over Union[21] (IOU) criteria at a 10% threshold, facilitated by a greedy matching assignment.

For the sake of clarity, our summarized results encompass the Dice score (prior to instance segmentation), the F1 score to evaluate the model's PRL detection capability, and the Dice Similarity Coefficient per True Positive (DSC/TP) for accurately segmenting true positives (TPs). These detection metrics are presented as subject-level averages.

4. RESULTS

Results shown in Table 1 revealed several key observations concerning the performance of different models in the detection and segmentation of PRL.

Firstly regarding models trained without undergoing any application of a mask, models trained on a single modality generally exhibited worse performance than those using multiple modalities, both for detection (avg F1 of 0.221 v. 0.336) and for segmentation tasks (avg. Dice of 0.152 v. 0.367). Notably, the FLAIR-based models showed relatively better segmentation performance compared to other single-modality approaches (avg. Dice of 0.369 v. 0.09). The magnitude-based models, however, displayed promising detection scores (F1=0.375) even though they had a low segmentation score. Interestingly, the *FLAIR+Phase* model emerged as the most successful single-modality approach in the absence of preprocessing (F1=0.456, Dice=0.377), aligning with previous findings from RimNet.[7]

Secondly, the application of a WM mask to the images tended to have a negative impact on overall results (avg. F1 diff.=-0.092; avg. Dice diff.=-0.71), whereas applying a lesion mask showed general improvements in detection (avg. F1 diff.=+0.115) and segmentation metrics (avg. Dice diff.=+0.017). The application of a lesion mask on

any input that included the FLAIR, however, tended to worsen the results (avg. Dice diff.= -0.094), suggesting that the FLAIR alone could potentially be sufficient to capture lesion localization accurately. Additionally, for segmentation metrics, models that utilized FLAIR as an input demonstrated the highest performance (avg. Dice 0.297 v. 0.140). This aligns with the fact that FLAIR sequences are particularly adept at highlighting lesions, and that annotators primarily relied on this modality to adjust their delineations.

Two models, *FLAIR+Phase* and *(FLAIR+MPRAGE)*Lesion Mask*, outperform the others in terms of the trade-off between detection and segmentation, both achieving a F1 score of 0.456, the highest among almost all models, and a Dice score of 0.377, better than all other models. Since the *FLAIR+Phase* model yielded a better DSC/TP, we applied this model to the test set and obtained a **F1 score of 0.443, a Dice score of 0.191, and a DSC/TP of 0.68**. A comparison of this model's prediction against the ground truth is shown on Figure 1.

5. DISCUSSION

The evaluation of various models and preprocessing strategies provided valuable insights into the challenges and potential solutions for automating PRL detection and segmentation. Firstly, we observed that models trained with a single modality as input generally exhibited suboptimal performance, with only FLAIR and Magnitude demonstrating relatively better results in detection. The low segmentation scores for magnitude-based models however contrast with those of FLAIR-based models. This can either be explained by the fact that this modality does not bring sufficient spatial information for accurate lesion delineation, or because experts used the FLAIR as their main image to provide a ground thruth. To mitigate the influence of FLAIR on annotations and to enhance the model's emphasis on Phase-presented rims, a compelling approach could involve training the network with expert annotations that were established using T2*-EPI as a reference, rather than FLAIR. Such a strategy could avoid registration errors that might lead to incomplete rim presence in the ground truth segmentation. Nevertheless, these observations highlight that the FLAIR sequence remains useful for capturing PRL characteristics, even though the rim is not visible in this modality. Moreover, the addition of the Phase and/or MPRAGE modalities to the FLAIR input showed improvements in the detection task, suggesting that these modalities provide complementary information that enhances the network's ability to discriminate PRL from non-PRL and background. The application of hard attention mechanisms, involving the use of a WM mask, did not significantly benefit the detection task. However, when the model utilized the Phase as an input, the application of a dilated lesion mask led to a substantial improvement in performance. This finding indicates that models based on the Phase image benefit from additional information about the lesion's location, which is in line with previous work[7] as well as with the experts' annotation workflow.

In contrast, the results confirm that the task of segmenting PRL is challenging, as none of the tested models surpassed the Dice Score of 0.46 on the validation set obtained from perfectly segmenting all lesions, regardless of whether they were PRL or not. This highlights the need for better methods to help the network discriminate between PRL and non-PRL lesions. Furthermore, our results indicate that the detection approach based on aggregating by connected components may not be the most suitable method, and future research could explore direct instance segmentation or object detection for improved results. Notably, this study showcases the difficulty of the PRL segmentation task, even when compared to other MRI biomarkers in MS such as WM lesions, cortical lesions, or lesions presenting the central vein sign. Several factors contribute to this challenge, including the relatively small dataset size, the imperfections in annotations, registration errors, and MRI artifacts. Moreover, the faint appearance of the PRL rim in a lot of cases (see Figure 1 VII and VIII) further complicates the task.

In conclusion, our study represents a pioneering effort in directly segmenting PRL without relying on lesion splitting. While the results are preliminary, they confirm on the potential of leveraging the FLAIR and Phase sequences for accurate PRL detection and lay the foundation for future research in this unexplored area. By identifying the strengths and limitations of different modalities and preprocessing strategies, our work contributes to the ongoing exploration of novel imaging biomarkers and automated detection methods in the field of MS lesion analysis. As the discovery of PRL is relatively recent and not yet acquired in routine clinical practice, the dataset size and associated challenges contribute to the difficulty of this segmentation task. Nevertheless, the insights gained from our study offer valuable directions for future research and hold the potential to significantly impact MS diagnosis, prognosis, and patient care.

Input modalities	Mask	F1	Dice score	DSC/TP
FLAIR		0,248	**0.377**	**0.559 ±0.29**
FLAIR	Lesion	0,346	0.273	0.556 ±0.29
FLAIR	WM	0.219	0.279	0.484 ±0.24
Phase		~ 0	0.002	-
Phase	Lesion	**0,5**	0.264	0.443 ±0.26
Phase	WM	0,061	0.009	0.214 ±0.18
MPRAGE		0,221	0.104	0.385 ±0.22
MPRAGE	Lesion	0,360	0.289	0.403 ±0.26
MPRAGE	WM	0,127	0.08	0.26 ±0.28
Magnitude		0,375	0.166	0.461 ±0.2
Magnitude	Lesion	0,434	0.216	0.361 ±0.25
Magnitude	WM	0,115	0.138	0.371 ±0.31
FLAIR + Phase		0,456	**0.377**	0.558 ±0.29
FLAIR + Phase	Lesion	0.297	0.148	0.433 ±0.27
FLAIR + Phase	WM	0,347	0.316	0.552 ±0.3
FLAIR + MPRAGE		0,219	0.356	0.553 ±0.27
FLAIR + MPRAGE	Lesion	0,456	**0.377**	0.524 ±0.28
FLAIR + MPRAGE	WM	0,234	0.197	0.384 ±0.25
FLAIR + Phase + MPRAGE		0,334	0.367	0.552 ±0.26
FLAIR + Phase + MPRAGE	Lesion	0,271	0.304	0.479 ±0.27
FLAIR + Phase + MPRAGE	WM	0,107	0.2	0.501 ±0.29

Table 1: Overview of the detection (F1 score) and segmentation (Dice coefficient) metrics for all trained models on the validation set. The mask column indicates whether a mask has been applied as a preprocessing step or not, and if so which one (Lesion = Lesion mask, WM = White matter mask). Best scores for each metric are highlighted in bold. ("DSC/TP": Average Dice score per True Positive).

ACKNOWLEDGMENTS

M.W. is supported by TRAIL and the Walloon region.

REFERENCES

[1] Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., Mowry, E. M., Sorensen, P. S., Tintoré, M., Traboulsee, A. L., Trojano, M., Uitdehaag, B. M. J., Vukusic, S., Waubant, E., Weinshenker, B. G., Reingold, S. C., and Cohen, J. A., "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," **17**(2), 162–173.

[2] on behalf of the MAGNIMS study group, "MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients," **11**(10), 597–606.

[3] Solomon, A. J., Naismith, R. T., and Cross, A. H., "Misdiagnosis of multiple sclerosis: Impact of the 2017 McDonald criteria on clinical practice," **92**(1), 26–33. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Views & Reviews.

[4] La Rosa, F., Wynen, M., Al-Louzi, O., Beck, E. S., Huelnhagen, T., Maggi, P., Thiran, J.-P., Kober, T., Shinohara, R. T., Sati, P., Reich, D. S., Granziera, C., Absinta, M., and Cuadra, M. B., "Cortical lesions, central vein sign, and paramagnetic rim lesions in multiple sclerosis: emerging machine learning techniques and future avenues,"

[5] Martire, M. S., Moiola, L., Rocca, M. A., Filippi, M., and Absinta, M., "What is the potential of paramagnetic rim lesions as diagnostic indicators in multiple sclerosis?," **22**(10), 829–837. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14737175.2022.2143265.

[6] Tejani, A. S., Retson, T. A., Moy, L., and Cook, T. S., "Detecting common sources of AI bias: Questions to ask when procuring an AI solution," **307**(3), e230580. Publisher: Radiological Society of North America.

[7] Barquero, G., La Rosa, F., Kebiri, H., Lu, P.-J., Rahmanzadeh, R., Weigel, M., Fartaria, M. J., Kober, T., Théaudin, M., Du Pasquier, R., Sati, P., Reich, D. S., Absinta, M., Granziera, C., Maggi, P., and Bach Cuadra, M., "RimNet: A deep 3d multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis," **28**, 102412.

[8] Zhang, H., Nguyen, T. D., Zhang, J., Marcille, M., Spincemaille, P., Wang, Y., Gauthier, S. A., and Sweeney, E. M., "QSMRim-net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps," **34**, 102979.

[9] Lou, C., Sati, P., Absinta, M., Clark, K., Dworkin, J. D., Valcarcel, A. M., Schindler, M. K., Reich, D. S., Sweeney, E. M., and Shinohara, R. T., "Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3t susceptibility-based MR imaging," 102796.

[10] Dworkin, J. D., Linn, K. A., Oguz, I., Fleishman, G. M., Bakshi, R., Nair, G., Calabresi, P. A., Henry, R. G., Oh, J., Papinutto, N., Pelletier, D., Rooney, W., Stern, W., Sicotte, N. L., Reich, D. S., and Shinohara, R. T., "An automated statistical technique for counting distinct multiple sclerosis lesions," **39**(4), 626–633. Publisher: American Journal of Neuroradiology Section: ADULT BRAIN.

[11] Smith, S. M., "Fast robust automated brain extraction," **17**(3), 143–155.

[12] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., and Xu, D., "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images."

[13] Cerri, S., Puonti, O., Meier, D. S., Wuerfel, J., Mühlau, M., Siebner, H. R., and Van Leemput, K., "A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis," **225**, 117471.

[14] Fischl, B., "FreeSurfer," **62**(2), 774–781.

[15] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation,"

[16] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H., "nnU-net: a self-configuring method for deep learning-based biomedical image segmentation," **18**(2), 203–211. Number: 2 Publisher: Nature Publishing Group.

[17] Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M. B., Gales, M. J. F., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.-J., Molchanova, N., Nikitakis, A., Raina, V., La Rosa, F., Sivena, E., Tsarsitalidis, V., Tsompopoulou, E., and Volf, E., "Shifts 2.0: Extending the dataset of real distributional shifts."

[18] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," **10553**, 240–248.

[19] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection."

[20] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, A. T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Van Calster, B., Varoquaux, G., and Jäger, P. F., "Metrics reloaded: Recommendations for image analysis validation."

[21] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., and Kirillov, A., "Boundary IoU: Improving object-centric image segmentation evaluation."