Impact of preprocessing on semantic segmentation of oocyte images

Alessandro Biscontin[1[0009-0006-7918-7438]], Milos Ajcevic[2[0000-0001-8307-4360]], Stefania Luppi[1[0000-0001-7391-9956]], Roberta Bottega[1[0000-0001-7431-7543]], Agostino Accardo[2[0000-0001-5749-6308]], and Giuseppe Ricci[1[0000-0002-8031-1102]]

[1] Institute for Maternal and Child Health-IRCCS "Burlo Garofolo", Trieste, Italy
[2] Department of Engineering and Architecture, University of Trieste, Via Valerio 10, 34127, Trieste, Italy
alessandro.biscontin@burlo.trieste.it

Abstract. The selection process of oocytes in In Vitro Fertilization (IVF) is still an open problem within the embryology research field. In recent years, several studies have been conducted to understand how image processing can help clinicians make better decisions. Moreover, the recent rise of deep learning due to the enhancement of computing capabilities added new evaluation strategies to the automatic technique's portfolio. In this study, we investigate whether the preprocessing phase is needed to fulfill the task of semantic segmentation. Twenty-three different deep neural network models were trained on two distinct datasets, one obtained from the original images acquired from the camera and the other obtained by preprocessing the raw images. The models obtained by using preprocessed data outperformed the ones obtained on raw images (Median Macro F1 score 0.8460 vs. 0.8222, respectively $p<0.001$). Furthermore, our analysis revealed that these models outperformed their counterparts even on the unprocessed dataset (Median Macro F1 score 0.8438 vs. 0.8222, respectively $p<0.001$). This, although a preliminary finding, suggests that every semantic segmentation pipeline should consider incorporating a preprocessing stage before feeding the data into the neural networks.

Keywords: Oocyte selection, Microscopy, Image processing, Semantic segmentation, Deep learning

1 Introduction

Oocyte selection is one of the fundamental phases of the Intracytoplasmic Sperm Injection (ICSI) process. Its main goal is to choose the best oocyte to be fertilized, hence increasing the chances of getting good embryos, and decreasing the probability that the possible pregnancy could end in miscarriage [1]. Dealing with cells that can potentially become new life, oocyte evaluation must be carried out with non-invasive methods to minimize the applied stress [2]. Within this context, the first and most intuitive strategy that embryologists came up with was to search for a correlation between cell morphometric characteristics and their fertilization potential.

The examination of these features continues to be a subject of debate since a universally accepted evaluation method for assessing the morphological components of oocytes has yet to be established [3]. This situation stems from several factors, e.g. a lack of research focused on quantitatively assessing the observed characteristics and a scarcity of comprehensive, multicenter studies in this field.

Morphological analysis is performed when the oocyte reaches the metaphase II arrest, that is the moment immediately preceding the beginning of the ICSI procedure, therefore it's expected that the oocyte already gained full maturation, which means it must gain both nuclear and cytoplasmic maturity [3]. Typically, the evaluation approach used by the embryologists is to qualitatively assess some morphological features. However, this traditional morphological assessment is subjective, and even for the same observer, it is difficult to repeat it in a reproducible way. To reduce this variability and improve selection accuracy, new technical approaches need to be implemented in the evaluation process [4]. One of the ways in which this can be achieved is using image analysis techniques, thereby transitioning from a selection based on qualitative assessments to one more focused on quantitative evaluation. In addition to making this process less subject to intra- and inter-operator variability, another benefit is to automate the process. Before proceeding with the actual analysis of the features, a fundamental step is the segmentation of the image into various Regions of Interest (ROI). In recent years, there has been increasing research interest in the application of deep learning techniques for performing this task, thanks to the enormous potential that this field has demonstrated. Specifically, in the field of biomedical imaging, the U-Net architecture was developed, which forms the basis of most of the algorithms used for semantic segmentation developed. Several recent studies perform semantic segmentation of oocyte images based on variation of the traditional U-Net architecture, adopting this implementation by focusing on different areas [5–8]. In some of these works, the segmentation is followed by the introduction of a pipeline, which is designed to assess the oocyte's competence, based on the measured morphological features. However, what those studies do not examine comprehensively is the preprocessing phase. Firuzinia et al. [5] and Barucic et al. [7] perform brightness standardization and minmax normalization, respectively. The other works does not mention any form of image preprocessing, except for resizing. Since it has been proven in other fields that preprocessing plays a crucial role in enhancing performance and generalization capabilities of the models [9], we wanted to investigate what was the effect that this phase has on the performance and robustness of models for semantic segmentation of oocyte images. Therefore, the main objective of this preliminary study was to perform a comparison of architectures trained on two distinct datasets, one based on original images and the other employing a preprocessed dataset.

2 Materials and Methods

2.1 Data acquisition and dataset definition

265 images of human oocytes in metaphase II, collected from 45 subjects undergoing the ICSI process at the Institute for Maternal and Child Health - IRCCS “Burlo Garofolo” in Trieste, have been gathered for dataset creation. The images were captured using an inverted microscope (Nikon Eclipse Ti) equipped with a CCD camera (QICAM Fast 1394 Cooled Digital Camera) at 200x magnification. The captured images are 1392 x 1040 pixels in size, for three 8-bit channels each. Manual segmentation was employed to label the images within the different ROIs. The segmented regions were cytoplasm, first polar body, perivitelline space, zona pellucida, cumulus/corona cells, and a sixth region comprising the background and any other spurious object. Regarding the cumulus/corona cells, segmentation included them if they were positioned in the foreground within the image's field of view, while cells out of the focal plane were ignored.

This study’s main goal is to assess the impact of preprocessing the dataset on training a specific model. To this end, two distinct datasets were generated, one with original images (OD), and the other with preprocessed images (PD).

2.2 Preprocessing

To generate PD, raw images underwent first the exposure correction and standardization, in which the gray levels were adjusted to align with the average gray level of all image backgrounds. For each image, the average gray level of a 75 x 75 pixels background patch was determined. The exposure-standardized image was then produced by adjusting the image's original values based on the patch's average and the dataset-wide patch average. Additionally, a sharpening filter was applied to enhance image contrast.

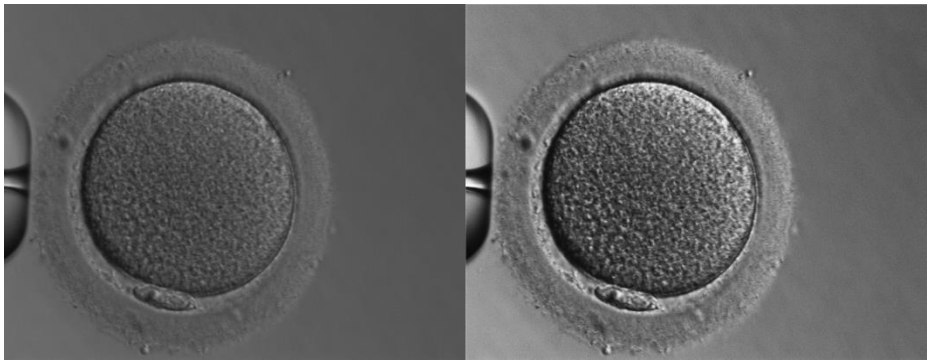

Fig. 1. Example of preprocessed image: on the left the photograph as it has been taken by the camera; on the right the exposure corrected and contrast enhanced image.

The OD images were kept in its original state, without any exposure correction or contrast enhancement. Both PD and OD were then resized to 256 x 256 pixels across three channels, optimizing the image size for neural network processing and ensuring detailed feature capture. The final preprocessing step involved normalizing the image values from the integer range [0, 255] to the floating-point range [0, 1].

2.3 Data augmentation

To reduce the risk of overfitting in trained models, a data augmentation phase was introduced for each dataset, applying various transformations to the images and their associated segmentation masks. This included applying random rotations between -45° and +45°, random horizontal and vertical shifts ranging from -26px to +26px, random zoom with scaling factors between -0.5x and 0.5x, and both vertical and horizontal flips. These augmentations enhance the robustness and generalizability of the models by providing them with a more diverse set of training examples.

2.4 U-Net architecture and variations

Introduced in 2015, the U-Net network quickly established itself in semantic segmentation with high accuracy, outperforming existing techniques. Its architecture features an encoder and a decoder with unique skip connections between corresponding blocks [10]. Building on this foundation, new variations have emerged: U-Net++ with nested convolutional blocks in the skip connections [11], Huang's architecture with full-scale skip connections [12], and Oktay's Attention U-Net with attention gates [13]. Other versions include the R2 U-Net, which employs residual and recursive structures [14], and the Transformer U-Net, which incorporates a transformer to better capture the global context [15].

2.5 Training

During the training process both datasets were divided into a training set (70% of the total images), a validation set (15%), and a test set (15%). For a more rigorous comparison of model performance, the training, validation, and test sets generated from OD and PD were meticulously constructed using identical acquired images in their respective raw and preprocessed format.

All deep neural network architectures presented in the previous paragraph have been trained. U-Nets variants were trained both from scratch and with backbones composed of a fully convolutional neural network pre-trained on the ImageNet dataset. The backbones used were VGG16, VGG19, ResNet50, and ResNet50V2. U-Net and Attention U-Net were trained both from scratch and with a backbone, U-Net++, UNet3+, and Transformer U-Net were trained only with the pre-trained backbones and R2-UNet were trained only from scratch.

ADAM with a learning rate of 0.001 was chosen as the optimization algorithm. Given the multi-class segmentation with unbalanced classes, weighted categorical

cross-entropy was used as the loss function to assign greater importance to errors in less represented classes:

$$wCCE = -\frac{\sum_{i=1}^{MxN}\sum_{c=1}^{C} w^c \cdot y_i^c \cdot \log(\hat{y}_i^c)}{M\ x\ N} \quad (1)$$

Where w^c represent to the weight assigned to the class c, M and N indicate the height and width of the image, y_i^c is the is the ground truth label for the i-th pixel, and $\hat{y}_i^c$ is the predicted label for the i-th pixel.

Specifically, classes corresponding to pixels in the cumulus/corona cells and first polar body areas were assigned a weight three times greater than that assigned to each of the other classes. The learning rate was reduced by a factor of 0.75 whenever the loss function did not decrease by more than 6 epochs from its lowest value. A maximum of 500 training epochs was set, which was never reached in any training session as it was stopped when the loss function did not improve for more than 20 epochs.

The entire training and validation sets were trained in batches, each containing 8 images. The best models are identified by selecting those with the lowest validation loss, and thus were used in the test phase.

2.6 Evaluation metrics

At the end of the training the models were evaluated on an independent test set. A confusion matrix was generated, detailing the correlation between the actual and predicted labels for each pixel across all images in the dataset. Therefore, for each class were extracted the True Positive TP, True Negative TN, False Positive FP, and False Negative FN values and then some performance metrics were evaluated.

For each class were assessed different metrics:

- Precision, to state the total number of pixels correctly segmented as class c:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (2)$$

- Recall, to indicate the total number of pixels of class c correctly individuated:

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (3)$$

- Dice similarity coefficient, as an overall accuracy metric for the class c:

$$DSC_c = 2\frac{TP_c}{2TP_c + FN_c + FP_c} \quad (4)$$

To describe the model's complete performance the Macro F1 score was then calculated. Since, for all the models some classes presented a conspicuous difference between precision and recall values, this metric was assessed as the mean value of all the Dice similarity coefficient [16]:

- Macro F1 score:

$$Macro\ F1\ = \frac{1}{C}\sum_{c=1}^{C} DSC_c \quad (5)$$

Where C is the total number of classes taken into consideration. The Macro F1 is computed only over the area useful to the analysis, not considering the performance on the background.

The Wilcoxon signed rank test was performed to assess the differences between performance metrics obtained on two different datasets. A p-value < 0.05 was considered statistically significant.

3 Results

Twenty-three different U-Nets were trained separately on PD and OD yielding a total of 46 models. Median and IQR of Macro F1 score were 0.8460 (0.8313-0.8531) for the models trained utilizing PD, and 0.8222 (0.7943-0.8378) for those trained utilizing OD. The best model obtained from the training was the U-Net model using ResNet50 as backbone, either among models trained with preprocessed image and among models trained with original images. For those two models, Macro F1 were 0.8665 and 0.8499, respectively. All their evaluated metrics are presented in Table 1.

Table 1. Precision, Recall and Dice Similarity Score for the ROIs of the two best models (with and without preprocessing).

Network	ROI	Precision	Recall	DSC
U-Net – ResNet50 backbone (trained with preprocessed dataset)	Cytoplasm	0.9940	0.9763	0.9851
	Perivitelline Space	0.8003	0.8916	0.8435
	Zona Pellucida	0.9462	0.8932	0.9190
	First Polar Body	0.6796	0.8954	0.7727
	Cumulus/Corona Cells	0.7756	0.8520	0.8120
U-Net – ResNet50 backbone (trained with original images)	Cytoplasm	0.9905	0.9844	0.9874
	Perivitelline Space	0.7749	0.8806	0.8244
	Zona Pellucida	0.9545	0.8650	0.9075
	First Polar Body	0.6252	0.9185	0.7440
	Cumulus/Corona Cells	0.7035	0.8910	0.7862

In Figure 2, Macro F1 scores of models obtained on preprocessed data were plotted against those obtained using original images. It can be observed that the models trained on PD performs better than those trained on OD, except for just one model. Indeed, the Wilcoxon signed-rank test showed that difference Macro F1 scores were significantly higher in PD-based models ($p < 0.001$). The median and IQR of the difference between Macro F1 scores were 0.02 (0.0125-0.0379).

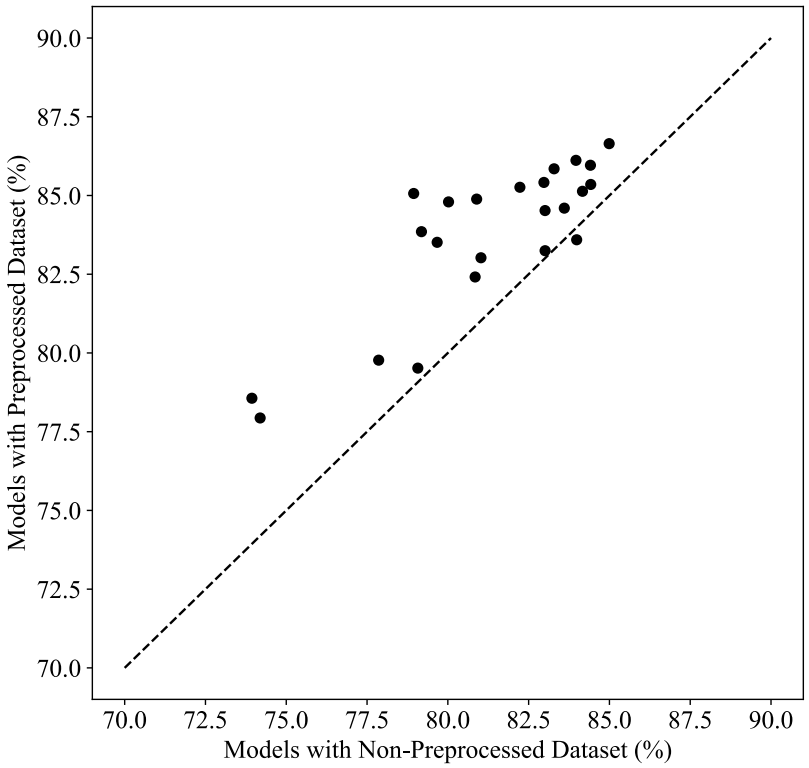

Fig. 2. Macro F1 score comparison: The x-axis represents scores for models trained on OD, and the y-axis illustrates scores for models trained on PD.

In addition, we also compared the behavior of the models on the set composed by original images. Again, models trained on the PD outperform those trained on the OD, as it can be seen in Figure 3. Wilcoxon signed-rank test was statistically significant ($p < 0.001$). Here, the median and IQR of difference between Macro F1 scores were 0.0182(0.0085-0.0324).

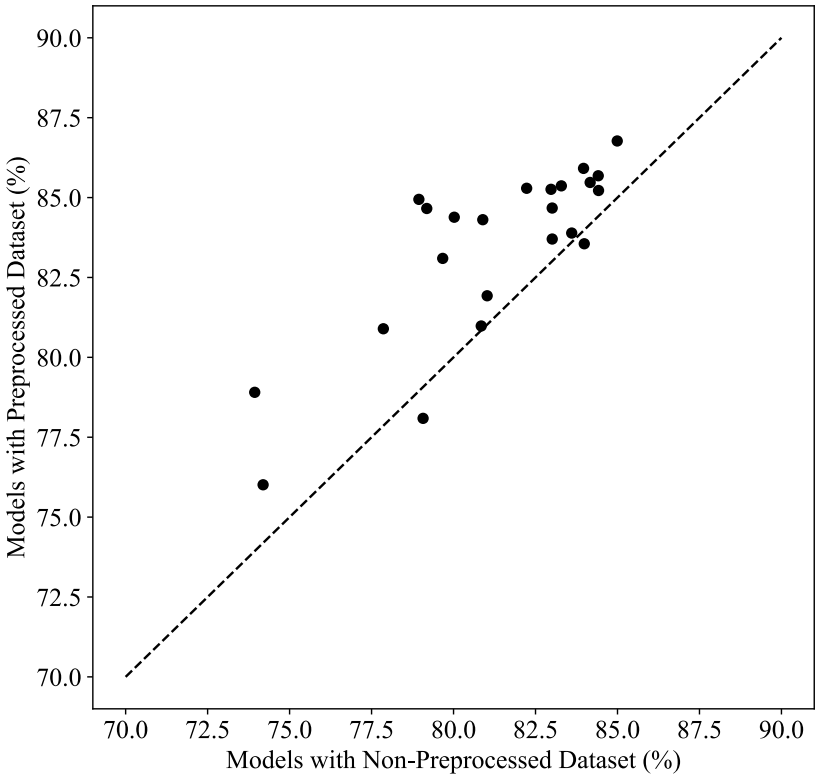

Fig. 3. Macro F1 score comparison for models trained on the test set from OD: on the x-axis the Macro F1 scores for the models trained on OD; on the y-axis the Macro F1 scores for the models trained on PD.

4 Discussion

In recent years, increasing interest has been shown by researchers in the field of semantic segmentation of biomedical images. In the presented study, we investigated how the preprocessing phase influence the performance of deep learning models developed for the semantic segmentation task. The results showed that using preprocessed images during this phase contributes to the development of more accurate models, with respect the metric considered. Moreover, those models perform better than OD-based models even if applicated on the test set composed by the original images, demonstrating an enhanced robustness.

Other works extended the application of this technique to human oocyte images, aiming to identify the optimal model for this task [6], whether it's better to make a multiclass segmentation or a binary segmentation [5] or to assess how it works into a pipeline of phenotyping oocytes [7, 8]. In our research we decided to segment 5 areas, which is a tradeoff between the approaches used in the other works presented in literature. This for two main reasons: first, the choice was made to have the minimum

number of areas so that, in the future, it will be possible to independently analyze every ROI that has been demonstrated to be clinically relevant; second, we wanted to not delve too deeply into the possible phenotypic variations of the areas, avoiding the risk of having insufficient data to achieve acceptable results for each class.

Furthermore, a comparison between the performance of our best model and the models proposed by Firuzinia et al. [5] and Targosz et al. [6], on a three areas semantic segmentation job (cytoplasm, perivitelline space and zona pellucida), is presented on Table 2.

Table 2. Comparison between best proposed model and other models presented in literature, according to Precision, Recall and Dice Similarity Score of the cytoplasm, perivitelline space and zona pellucida.

ROI	Model	Precision	Recall	DSC
Cytoplasm	Proposed	0.9940	0.9763	0.9851
	Firuzinia et al. [5]	0.9872	0.9897	0.9884
	Targosz et al. [6]	0.9838	0.9861	0.9849
Perivitelline Space	Proposed	0.8003	0.8916	0.8435
	Firuzinia et al. [5]	0.8981	0.9017	0.8999
	Targosz et al. [6]	0.8326	0.8706	0.8501
Zona Pellucida	Proposed	0.9462	0.8932	0.9190
	Firuzinia et al. [5]	0.9282	0.9408	0.9345
	Targosz et al. [6]	0.9127	0.9370	0.9235

The performance obtained from the proposed method is directly comparable with those presented in the literature, providing an architecture that could be considered in future for application in the clinical practice.

While this study has provided valuable insights in understanding the importance of preprocessing phase, it is important to acknowledge its limitations. Although the number of models considered was sufficient to produce an accurate analysis, not all the semantic segmentation architectures known in literature have been explored. Moreover, the restricted number of images and the scarce variability of the samples used in the training phase are other limiting factors for this deep learning application, causing the possible lack of generalization of the produced models.

Therefore, these results should be confirmed in a large clinical study with a wider dataset – possibly with images acquired from different cameras or oocytes coming from different IVF centers – and with more semantic segmentation models.

In conclusion, in this study we preliminarily assessed that preprocessing stage gives an added value of the segmentation of the oocytes and the models produced on preprocessed data generalize better and are more robust. Furthermore, the proposed method showed a similar performance metrics compared to state-of-art literature.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Ozturk, S.: Selection of competent oocytes by morphological criteria for assisted reproductive technologies. Molecular Reproduction and Development. 87, 1021–1036 (2020). https://doi.org/10.1002/mrd.23420.
2. Sciorio, R., Miranian, D., Smith, G.D.: Non-invasive oocyte quality assessment. Biology of Reproduction. 106, 274–290 (2022). https://doi.org/10.1093/biolre/ioac009.
3. Wang, Q., Sun, Q.-Y.: Evaluation of oocyte quality: morphological, cellular and molecular predictors. Reprod. Fertil. Dev. 19, 1 (2007). https://doi.org/10.1071/RD06103.
4. Palay, P., Fathi, D., Fathi, R.: Oocyte quality evaluation: a review of engineering approaches toward clinical challenges. Biology of Reproduction. 108, 393–407 (2023). https://doi.org/10.1093/biolre/ioac219.
5. Firuzinia, S., Afzali, S.M., Ghasemian, F., Mirroshandel, S.A.: A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. Computer Methods and Programs in Biomedicine. 201, 105946 (2021). https://doi.org/10.1016/j.cmpb.2021.105946.
6. Targosz, A., Przystałka, P., Wiaderkiewicz, R., Mrugacz, G.: Semantic segmentation of human oocyte images using deep neural networks. BioMed Eng OnLine. 20, 40 (2021). https://doi.org/10.1186/s12938-021-00864-w.
7. Baručić, D., Kybic, J., Teplá, O., Topurko, Z., Kratochvílová, I.: Automatic evaluation of human oocyte developmental potential from microscopy images. In: 17th International Symposium on Medical Information Processing and Analysis. p. 2 (2021). https://doi.org/10.1117/12.2604010.
8. Letort, G., Eichmuller, A., Da Silva, C., Nikalayevich, E., Crozet, F., Salle, J., Minc, N., Labrune, E., Wolf, J.-P., Terret, M.-E., Verlhac, M.-H.: An interpretable and versatile machine learning approach for oocyte phenotyping. Journal of Cell Science. jcs.260281 (2022). https://doi.org/10.1242/jcs.260281.
9. Li, A., Winebrake, J.P., Kovacs, K.: Facilitating deep learning through preprocessing of optical coherence tomography images. BMC Ophthalmology. 23, 158 (2023). https://doi.org/10.1186/s12886-023-02916-2.
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., and Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28.
11. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning

for Clinical Decision Support. pp. 3–11. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1.
12. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J.: UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059. IEEE, Barcelona, Spain (2020). https://doi.org/10.1109/ICASSP40776.2020.9053405.
13. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas, http://arxiv.org/abs/1804.03999, (2018).
14. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation, http://arxiv.org/abs/1802.06955, (2018). https://doi.org/10.48550/arXiv.1802.06955.
15. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, http://arxiv.org/abs/2102.04306, (2021).
16. Opitz, J., Burst, S.: Macro F1 and Macro F1, http://arxiv.org/abs/1911.03347, (2021).