Development of a Deep-learning Pipeline to Recognize and Characterize Macrophages in Colorectal Liver Metastasis

Pierandrea Cancian
Humanitas Research Hospital
Nina Cortese
Humanitas Research Hospital
Matteo Donadon
Humanitas University
Marco Di Maio
Humanitas Research Hospital
Cristiana Soldani
Humanitas Research Hospital
Federica Marchesi
University of Milan
Victor Savevski
Humanitas Research Hospital
Marco Domenico Santambrogio
Polytechnic University of Milan
Luca Cerina
Polytechnic University of Milan
Guido Torzilli
Humanitas University
Alberto Mantovani
Humanitas Research Hospital
Luigi Terracciano
Humanitas University
Massimo Roncalli
Humanitas University
Luca Di Tommaso (luca.di_tommaso@hunimed.eu)
Humanitas University

Research Article

Keywords: tumor microenvironment (TME), cancers , colo-rectal liver metastases (CLM), tumor associated macrophages (TAM)

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-124192/v1

Development of a deep-learning pipeline to recognize and characterize macrophages in colo-rectal liver metastasis.

Pierandrea Cancian[1,2], Nina Cortese[3], Matteo Donadon[4,6,], Marco Di Maio[2], Cristiana Soldani[5], Federica Marchesi[3,7,*], Victor Savevski[1], Marco Domenico Santambrogio[8], Luca Cerina[8], Guido Torzilli[4,6], Alberto Mantovani[3,6], Luigi Terracciano[1,6], Massimo Roncalli[1,6], Luca Di Tommaso[1,6,*].*

[1]Artificial Intelligence Center, [2]Department of Pathology, [3]Department of Immunology and Inflammation, [4]Department of Hepatobiliary and General Surgery and [5]Hepatobiliary Immunopathology Unit, Humanitas Clinical and Research Center - IRCCS, via Manzoni 56, 20089 Rozzano (Mi) – Italy; [6]Department of Biomedical Science, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele – Milan, Italy; [7]Department of Biotechnology and Translational Medicine, University of Milan, Milan, Italy; [8]Dipartimento Informazione, Elettronica e Bioingegneria, Politecnico di Milano, Milan, Italy

Corresponding authors:
Luca Di Tommaso,
Department of Pathology, Humanitas Clinical and Research Center
Via Manzoni 56, 20089 Rozzano (Mi), Italy. Tel: +390282244715 Fax: +390282244791
Email: luca.di_tommaso@hunimed.eu

Federica Marchesi,
Department of Immunology and Inflammation
Humanitas Clinical and Research Center
Via Rita Levi Montalcini 4, 20090 Pieve Emanuele (MI) Italy.
Email: federica.marchesi@humanitasresearch.it

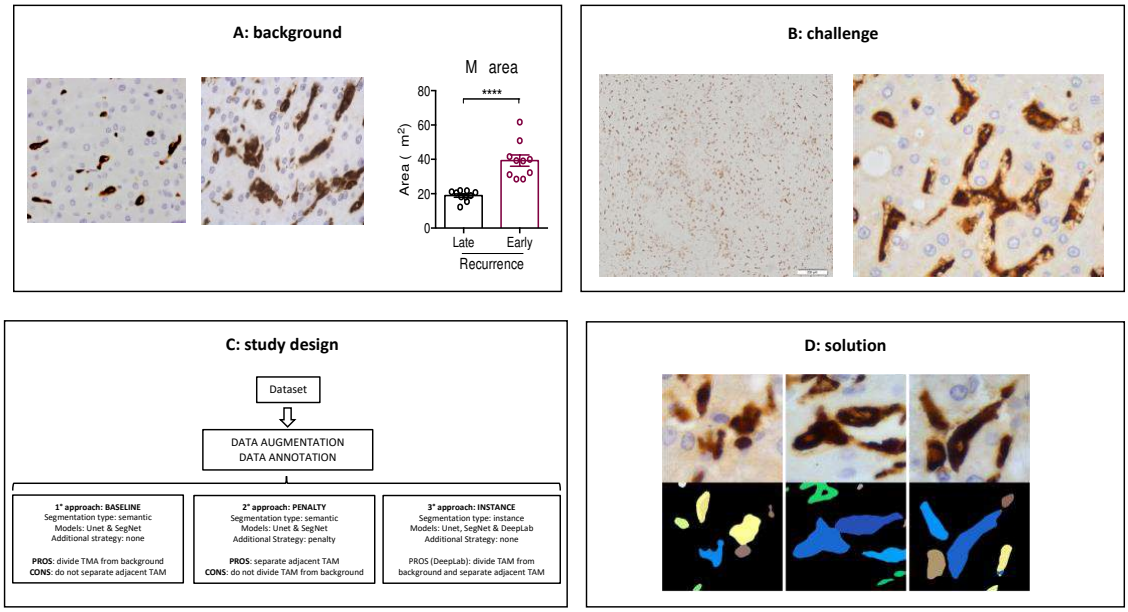
Matteo Donadon,
Department of Hepatobiliary and General Surgery, Humanitas Clinical and Research Center
Via Manzoni 56, 20089 Rozzano (Mi), Italy.
Email: matteo.donadon@hunimed.eu

ABSTRACT

Quantitative analysis of tumor microenvironment (TME) provides prognostic and predictive information in several human cancers but, with few exceptions, it is not performed in the daily clinical practice being time-consuming. We recently showed that the morphology of tumor associated macrophages (TAM) correlates with outcome in patients with colo-rectal liver metastases (CLM). However, as for other TME components, recognizing and characterizing hundreds of TAM in a single histopathological slide is unfeasible. To fasten this process, we explored a deep-learning based solution.

We tested three convolutional neural networks (CNN), Unet, SegNet and DeepLab-v3, and compared their results according to IoU (intersection over union), a metric describing the similarity between what CNN predicts as TAM and the ground truth and SBD (symmetric best dice), which indicates the ability of CNN to separate different TAMs. Unet and SegNet showed intrinsic limitations in discriminating single TAMs (highest SBD 61.34 ± 2.21), whereas DeepLab-v3 accurately recognized TAM from the background [IoU (89.13 ± 3.85)] and separated different TAM [SBD (79.00 ± 3.72)]. This deep-learning pipeline to recognize TAMs in digital slides, will allow the characterization of TAM-related metrics in the daily clinical practice, allowing the implementation of prognostic tools.

Visual abstract of the study.



The background. We recently proved that in human colorectal cancer, the presence of small and large tumor-associated macrophages (TAMs), is associated with different outcomes. **The challenge**. To translate this biological data into a robust clinical marker means to identify all TAMs in a single slide, which largely outnumber hundreds of cells, and then evaluate the area of each of them, a task unfeasible in the routine pathology workout. **Study design**. With the aim to develop a deep-learning pipeline to tackle this challenge, we selected, trained and tested three different approaches. **The solution**. The deep-learning pipeline based on the DeepLab-v3 architecture and semantic segmentation technique warrants the separation of TAMs from the background and the identification of single TAMs: this will easily allow the evaluation of their area.

INTRODUCTION

During the last two decades, technological advances have transformed histopathological glass slides into high-resolution digital slides [1]. The availability of digital slides, in turn, has allowed the introduction and application of image analysis methods to histopathology, as previously happened to radiology. Image analysis methods enables recognizing, differentiating, and quantifying images and finally sustains the development of Computer-Aided Diagnosis (CAD) tools. Once applied to images, these instruments support the diagnostic process, by highlighting a variety of aspect of interest. Among image analysis methods, great relevance has been progressively acquired by deep learning [2]. These latter models are characterized by an input layer (image data), hidden layers and an output layer (predictions): a representation of human neural architecture known as artificial neural networks [3]. A specific neural network architecture known as convolutional neural networks (CNN) is the standard for image recognition [4,5]. CNN are essentially made by several filters, used to extract the high number of information available within images and aggregate them into a lower amount, still relevant to complete the assigned duty. In the field of histopathology, CNN have already been used for several tasks: from the detection of a simple object as a mitotic figure [6], trough the classification of prostate cancer grading [7], to the intriguing potential to pick up, from a simple H/E staining, information regarding the prognosis [8], the response to treatment [9] or even the presence of molecular alterations [10].

Malignant tumors are composed of a heterogeneous population of cancer cells, admixed with a variety of host cells and secreted molecules, namely the tumor microenvironment (TME), whose dynamic interactions determine whether the tumor is eradicated or progresses, in a Darwinian-type evolutionary process [11]. Efforts aimed at implementing deep-learning solutions have been convincingly made also in the onco-immunology field, where quantitative analysis of immune cells and TME components has produced relevant results in terms of identification of prognostic biomarkers and better patient profiling [12-14]. The possibility to adopt a CAD method for the evaluation of T cells has been robustly tested and validated in colo-rectal cancer [12,15]. As to macrophages, plenty of which populate the tumor microenvironment [16-21], there is no clear association with clinical outcomes across cancers [22], possibly due to their profound heterogeneity in terms of polarization, functions, and tissue localization.

We have recently shown that tumor associated macrophage (TAM) morphology is associated with distinct transcriptomic profiles and clinical outcomes in human colorectal liver metastases (CLM) [21]. In particular, when we separated TAM according to their area, we observed that small (S) and large (L) TAM correlated with a 5-year disease-free survival rate of 27.8% and 0.2% respectively ($P<0.0001$). This was associated to different molecular profiles of small and large populations,

particularly in their lipid metabolism and phagocytic repertoire. However, despite the fact that the prognostically negative TAM can be identified under the microscope, being characterized by larger size, irregular borders and foamy cytoplasm, the assessment and categorisation of TAM morphology is unfeasible in the daily clinical practice, since they are admixed with hundreds of other similar macrophages, with distinct transcriptional profile or prognostic values. Overall, their recognition and proper characterization would result in being time consuming and an impossible task for clinical purposes. Based on these premises, we aimed at developing a deep-learning pipeline able to systematically recognize all TAM in a CLM specimen. Future studies aimed at evaluating TAM-related features as prognostic or predictive markers would benefit of this digital tool. Once validated in large cohorts, this tool could be integrated in the routine assessment of human CLM and aid in the histopathological report, with consequent translational impact.

RESULTS and DISCUSSION

Overview of the strategy adopted to choose CNN models for TAM segmentation.

The goal of this study was the development of a deep-learning pipeline for the segmentation, i.e. the proper identification at pixel level, of single TAM in digital slides of human liver parenchyma. The ultimate output was a color mask, dividing foreground (TAM) from background (hepatocytes, cholangiocytes, other components) as well as the separation of single TAM.

To this aim, we selected, trained and tested three different CNN models: the Unet [23], the SegNet [24] and the DeepLab-v3 [25] with a MobileNet-v2 [26] backbone. The UNet and the SegNet were used at the beginning of the study to establish a baseline (**Table 1**), because they had been already employed for segmentation tasks in histopathology [3,27]. This approach showed good results in separating TAM from the background, but failed to accurately separate adjacent TAM. To overcome this problem, we decided to try a different strategy. Without changing the UNet and the SegNet architecture, we assigned different weights to different pixels contained in regions between one TAM and another; nonetheless, this penalty approach was not sufficient to achieve separation of adjacent TAM (**Table 1** and **Figure 1)**. Considering this limit, we decided to shift from semantic to instance segmentation. Briefly, semantic segmentation assigns to each pixel a class, but it considers multiple objects of the same class as a single entity. By contrast, instance segmentation warrants identifying not only the class but also the individual object (instance) each pixel belongs to. We also introduced a more powerful architecture, the DeepLab-v3 (**Table 1** and **Figure 2)**.

The results are discussed in the specific sections.

Table 1. Quality of the separation mask (foreground *vs.* background) in the baseline, penalty and instance approaches.

Model	*IoU mean*	*IoU standard deviation*	*SBD mean*	*SBD standard deviation*
UNet baseline	82.34	2.59	39.87	2.90
SegNet baseline	82.13	1.59	40.59	4.80
UNet penalty	59.62	3.15	61.34	2.21
SegNet penalty	54.43	3.87	53.90	3.47
UNet instance	75.41	5.30	37.15	5.30
SegNet instance	79.98	2.15	60.79	5.75
DeepLab instance	89.13	3.85	79.00	3.72

***Legend**. IoU: intersection over union; SBD: symmetric best dice.*

Metrics and Baseline results

We evaluated two metrics to assess the quality of the separation mask, intersection over union (IoU) and symmetric best dice (SBD) (see Methods for detailed description). Briefly, IoU describes the similarity between the predicted segmentation mask (what the model predicts as TAM) and the ground truth mask (what was originally annotated as TAM), whereas SBD describes the ability to separate different instances (TAMs). In the baseline analysis, UNet and SegNet showed high IoU values (IoU = 82.34±2.59 and 82.13±1.59, respectively), highlighting their ability to properly localize the pixels belonging to TAM. At the same time, low SBD values (SBD = 39.87±2.90 and 40.59±4.80, respectively) suggested the inability to separate adjacent TAM (**Figure 1**).

Taking into consideration that the IoU and SBD have very similar underlying mechanics, the divergent results suggest that the CNN models work properly only in trivial cases, blindly assigning dark patches to the foreground. Most likely, the models were not able to create meaningful high-level features representing the single cells, thus leveraging only low level features such as color. This also indicates intrinsic weakness of the models we adopted for the specific task of this study. In keeping with this, it has been reported that UNet can hardly distinguish touching or overlapping cells, as well as some confusing background areas [28]. However, it is also possible that semantic segmentation *per se* is not the most appropriate approach for our task. Indeed, the semantic technique treats multiple objects of the same category as a single entity: its goal is to recognize the correct class, not to separate properly each instance belonging to the same class.

Penalty approach

Considering the limits raised by the baseline analysis, we decided to work towards improving the ability of the models in separating different TAM. To this aim, we adopted the strategy of assigning a penalty to the misclassified pixels in between TAM, an approach outlined in the first UNet publication [23]. As expected, we observed an important increase of SBD for both models (SBD = 61.34±2.21 and 53.90±3.47 for UNet and SegNet respectively), with an increase in the capacity to separate adjacent TAM, as shown in **Figure 1**. Nonetheless, the absolute values of SBD (**Table 1**) were in keeping with a persistent limit on this specific assignment. By contrast, the other metric, IoU, decreased significantly (IoU = 59.62±3.15 and 54.43±3.87 for UNet and SegNet respectively) because the models showed very prominent artifacts in the segmentation mask. In addition, the penalty strategy had non-trivial overhead in terms of development time. Computing the weights of the ground truth images is time consuming and requires considerable tweaking of the parameters, such as finding an effective value for the penalty or deciding which pixels should be penalized. A further aspect to be taken into account is that imposing the background label in between foreground

objects is not a natural way to solve the problem, for humans and models alike, but rather a workaround to coerce the model into behaving as it is expected to work. This is particularly clear, for example, when considering that a digitized histological slide is a 2D representation of a 3D environment, and this dimensionality reduction often leads to an actual overlap in the 2D projection of cells closely located in the 3D specimen. The background strip artificially imposed by the penalty approach is merely an expedient that does not reflect the real architecture of the 3D environment.

Figure 1

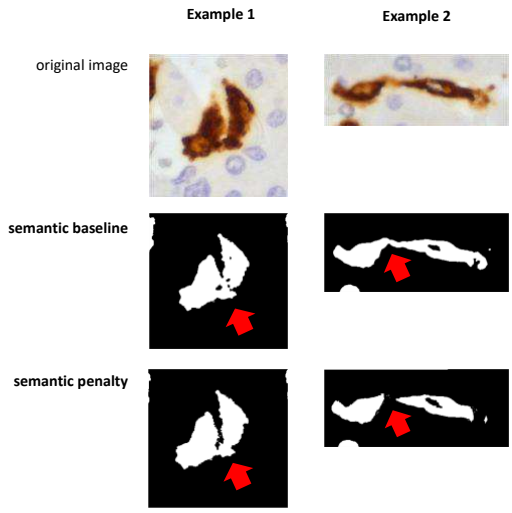


Figure 1. TAM segmentation: semantic approach. *The figure compares the histopathological image and the output of CNN model (UNet in these examples) using a semantic approach. In the original image TAM are stained in brown by CD163 immunostaining and the surrounding hepatocytes (the background) are counterstained in blue; (nuclei darker than cytoplasm); in CNN output TAM are returned in white and the background in black. The* ***semantic baseline analysis*** *generated an optimal separation between TAM and the background but was not able to separate adjacent TAMs, as highlighted by red arrows. Adopting the* ***semantic penalty*** *approach fixed part of this specific problem: the two TAM shown on the right are properly characterized by this solution.*

Instance Segmentation approach

The results of the penalty-based approach further support the hypothesis that, despite the employment of a time-consuming strategy, the semantic approach can only attribute the pixel to the correct class (TAM) but not the single members of the class (TAMs). To address this aspect, in the third part of the study we shifted from a semantic to an instance segmentation approach. Specifically, we used the discriminative loss function proposed by De Brabandere et al [29] to transform semantic segmentation models into instance segmentation models. Moreover, we introduced, in addition to UNet and SegNet, the state-of-the-art CNN architecture, DeepLab-v3. The models output two different segmentation masks: semantic and instance. In this part of the study, the training was performed with Cross-Entropy

loss functions for the semantic output and the discriminative loss function for the instance output. Interestingly, for both UNet and SegNet, the metrics generated with the instance technique largely overlapped those of the semantic approach, confirming the intrinsic limitations of these models to separate TAMs. By contrast, the DeepLab-v3 model with instance segmentation showed a robust increase for both IoU (IoU = 89.13±3.85) and, in particular, for SBD (SBD = 79±3.72) (**Table 1**), allowing, as shown in **Figure 2**, the proper identification of adjacent TAM.

Figure 2

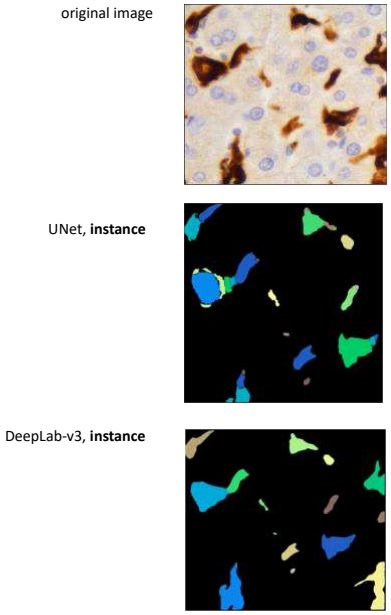


Figure 2. TAM segmentation: instance approach. *This figure compares the histopathological image and the outputs given by two models (UNet and DeepLab-v3) using instance segmentation. The output generated by Unet is characterized several instances (each characterized by a specific color) outnumbering true TAM. By contrast, Deep Lab-v3 model properly labelled each TAM as single instance, showing also a good performance to separate adjacent cells.*

Conclusion

In this study, we developed a deep-learning based pipeline that successfully recognizes and separates TAMs in digital slides of CLM. A variety of factors within the TME, in particular immune components, have been shown to robustly associate with clinical outcome and therapeutic responses [12-14,30]. Despite this, clinically feasible methods to quantitatively and reproducibly evaluate key immune elements are still missing [31]. Notable exception is the Immunoscore in human colo-rectal

cancer, a robust prognostic scoring system [12,15] evaluated on digital slides using a CAD tool [32] and recently introduced among the “Essential Criteria” in WHO classification of digestive system tumors [33]. Most of the remaining immune variables can be evaluated only with time-consuming, poorly-reproducible methods which make them unsuitable for practical purposes.

The pipeline that we presented here was developed to overcome limitations due to identification of macrophages in liver parenchyma. Progressive digitalization of histopathology slides and development of a deep learning approach allowed us to execute this process, i.e. identification of macrophages, faster. Indeed deep-learning solution can automate and significantly speed-up time-consuming procedures, including searching for a specific cellular type on digital histopathological slides [34]. Interestingly enough, our analyses showed the best results with DeepLab-v3 supported by the MobileNet-v2 backbone and the Atrous Spatial Pyramid Pooling (ASPP) module. Altogether these elements allowed the model to investigate at the deepest level the heterogeneity of TAM morphology (avoiding over-segmentation of larger cells) and the complexity of their spatial relationships (avoiding grouping of adjacent cells).

Open questions remain on the possibility to analyze regions of interest (ROI), compared to the whole section and to allow the operator or the program to select the ROI. The most effective strategy could be obtained by a balanced integration of the manual and automatic approaches.

In the near future, this pipeline will be tested as a CAD tool to perform quality check over manual annotation, to distinguish, for instance, S- from L-TAM. Other quantitative applications are under study. The ultimate output would be to introduce this pipeline in the routine workout of a pathological report, in order to complement and integrate the manual annotation

MATERIALS AND METHODS

Dataset.

This study is based on the dataset of images used in the paper by Donadon et al [21].

Data Augmentation.

To virtually enlarge the dataset we applied aggressive data augmentation [29,35]. This process applies any of the following transformation to the input data: rotation, scaling, shear, warp, color jitter, and changes to contrast, hue and exposure. In particular, morphological transformations are applied with a random factor in the (0.8, 1.3) interval, while the rotations can happen at any angle.

Data annotation.

Data annotation required drawing the contour of every single TAM in every single original histopathological image, a procedure performed under the strict supervision of an expert pathologist (LDT) using an open-source graphic-manipulation software. TAMs were then assigned to the "foreground" class while hepatocytes, cholangiocytes and other normal structures of liver parenchyma were grouped together in the "background" class. At the end of this process original histopathological pictures were coupled with black (background) and white (TAM) images which provide the ground truth, i.e. the solution to the problem. The CNN models returned similar images as output: black and white for semantic segmentation and color for instance segmentation.

Deep learning models.

The UNet [23] and the SegNet [24] represent the two different ends of the field-of-view trade-off. The former produces very sharp contours and very few artifacts, however, being a shallow network, its main limitation is the inability to valorize deeply the global information within the image. The latter is a deeper architecture with a more lightweight decoder path and smaller memory footprint.
The DeepLab-v3 architecture outperforms other simpler architectures [25] due to the MobileNet-v2 backbone [26] and the Atrous Spatial Pyramid Pooling (ASPP) module [36]. The former allows to restrict the number of parameters under evaluation. The latter allows to explicitly tackle the same objects appearing at different sizes, that confound simpler models in the task of instance segmentation leading to over-segmentation of larger cells. Briefly, this is performed by processing the features extracted by the backbone independently with atrous convolutions at different scales and merging the final results.

All the models were trained until the validation loss diverged from the training loss. We used the Adam optimizer with a learning rate of 1e-4, decaying when the loss stalled for a certain number of consecutive iterations [37]. We used the cross entropy loss at the pixel level for all the experiments and, in the case of instance segmentation, we added the discriminative loss function described in [29]. The models were fed randomly extracted 224x224 patches at 60x magnification.

Deep learning strategies.

Per Pixel Weights. We assigned different weights to different pixels when computing the loss during training; pixels with a higher weight contributed more to the final loss when misclassified; on the contrary those with lower weight incurred in less penalty. In particular, the pixels lying between TAMs that were closer than 10 pixels were associated with a penalty of about thrice as much, to force the model to recognize the single instance of every cell. We also found beneficial to reduce the penalty for all the pixels on the edge of the cells, because often the outlines of the cells were blurry and should not be considered a perfect label. The weights have been determined experimentally, running several trials until the results converged to a satisfying result.

Instance Segmentation. Instance segmentation is usually based on two different approaches: object-detection or pixel embedding. Object-detection based approaches showed the most promising results in the literature on most datasets but have some shortcomings that are particularly relevant in the field of histology: firstly, it is not suited for small datasets, and secondly, it struggles with string-like objects and cluttered structures, such as cells. Accordingly, we opted for the pixel embedding strategy, modifying a semantic segmentation model following the work by De Brabandere et al [28] with minor modifications. These authors proposed a new loss function that drives pixels belonging to the same instance in discrete clusters by embedding them in a high dimensional space. The single clusters are then isolated with a standard clustering algorithm at prediction time. In the original work the model's last convolutional layer is duplicated to produce two outputs, one for conventional semantic segmentation and one n-dimensional embedding space. In our work we found that duplicating the entire decoder branch of a DeepLab-v3 architecture led to the best results.

Metrics.

We use two different metrics for each experiment to describe the quality of the segmentation mask: the "Intersection over Union" (IoU), as a metric for semantic segmentation quality, and the "Symmetric Best Dice" (SBD**)** as a metric for instance segmentation quality.

The IoU is a statistic for measuring the similarity between the predicted segmentation mask and the ground truth mask; it is defined as the size of the *intersection* divided by the size of the *union* of the

masks. With P the predicted mask and T the ground truth mask, we can express this concept with the formula $P \cap T / P \cup T$.

The SBD quantifies how much the single instance overlaps with the most similar ones. It is an affirmed way of assessing the instance separation quality. To understand the formulation of this statistic we have to start from the concept of Dice Score, defined in a similar fashion to the IoU as: $2 * (P \cap T)/P + T$. The dice score tends to penalize less the worst performing cases than the IoU. Furthermore we can define the Best Dice for two segmentation mask L^a and L^b , with M and N instances respectively as:

$$\mathbf{BD}\left(L^a, L^b\right) = \frac{1}{M}\sum_{i=1}^{M} \max_{1\leq j\leq N} 2\frac{\left|L_i^a \cap L_j^b\right|}{|L_i^a| + \left|L_j^b\right|}.$$

In words: for each instance *a* in L^a find the instance *b* in L^b that maximizes the Dice Loss, then return the mean of all the computed partial results. The metric is not commutative with respect to L^a and L^b , as:

$$\mathbf{BD}\left(L^a, L^b\right) \neq \mathbf{BD}\left(L^b, L^a\right).$$

The SBD deals with this by taking the minimum of the two. It follows that:

$$\mathbf{SBD} = \min\left(BD(X,Y), BD(Y,X)\right).$$

REFERENCES

1. Griffin, J., and Treanor, D. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology.* **70**, 134-145 https://doi.org/10.1111/his.12993 (2017)
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature*. **521**, 436–444 https://doi.org/10.1038/nature14539 (2015).
3. Haque, I.R.I & Neubert, J. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked.* **18**, 100297 https://doi.org/10.1016/j.imu.2020.100297 (2020).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. **25**, 1097–1105 (2012).
5. Szegedy, C. et al. Going deeper with convolutions. https://arxiv.org/abs/1409.4842 (2014).
6. Wang, H., et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of medical imaging (Bellingham, Wash.)*. **1**, 034003 https://doi.org/10.1117/1.JMI.1.3.034003 (2014).
7. Bulten, W., et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology.* **21**, 233-241 https://doi.org/10.1016/S1470-2045(19)30739-9 (2020).
8. Yu, K. H., et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*. **7**, 12474 https://doi.org/10.1038/ncomms12474 (2016).
9. Coudray, N., et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine*. **24**, 1559–1567 https://doi.org/10.1038/s41591-018-0177-5 (2018).
10. Schmauch, B., et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature communications*. **11**, 3877 https://doi.org/10.1038/s41467-020-17678-4 (2020).
11. Angelova, M., et al. Evolution of Metastases in Space and Time under Immune Selection. *Cell*. **175**, 751–765 https://doi.org/10.1016/j.cell.2018.09.018 (2018).
12. Galon, J., et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science.* **313**, 1960–1964 https://doi.org/10.1126/science.1129139 (2006).

13. Laghi, L., et al. CD3+ cells at the invasive margin of deeply invading (pT3-T4) colorectal cancer and risk of post-surgical metastasis: a longitudinal study. *The Lancet. Oncology*. **10**, 877–884 https://doi.org/10.1016/S1470-2045(09)70186-X (2009).
14. Fridman, W. H., Zitvogel, L., Sautès-Fridman, C., & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nature Reviews Clinical Oncology*. **14**, 717–734 https://doi.org/10.1038/nrclinonc.2017.101 (2017).
15. Cortese, N., Carriero, R., Laghi, L., Mantovani, A., & Marchesi, F. Prognostic significance of tumor-associated macrophages: past, present and future. *Seminars in immunology*. 101408 https://doi.org/10.1016/j.smim.2020.101408 (2020).
16. Mantovani, A., Allavena, P., Sica, A., & Balkwill, F. Cancer-related inflammation. *Nature*. **454**, 436–444 https://doi.org/10.1038/nature07205 (2008).
17. Ruffell, B., Affara, N. I., & Coussens, L. M. Differential macrophage programming in the tumor microenvironment. *Trends in immunology*. **33**, 119–126 https://doi.org/10.1016/j.it.2011.12.001 (2012).
18. Murray, P. J., et al. Macrophage activation and polarization: nomenclature and experimental guidelines. *Immunity*. **41**, 14–20 https://doi.org/10.1016/j.immuni.2014.06.008 (2014).
19. Mantovani, A., Marchesi, F., Malesci, A., Laghi, L., & Allavena, P. Tumour-associated macrophages as treatment targets in oncology. *Nature Reviews Clinical Oncology.* **14**, 399–416 https://doi.org/10.1038/nrclinonc.2016.217 (2017).
20. DeNardo, D. G., & Ruffell, B. Macrophages as regulators of tumour immunity and immunotherapy. *Nature Reviews Immunology*. **19**, 369–382 https://doi.org/10.1038/s41577-019-0127-6 (2019).
21. Donadon, M., et al. Macrophage morphology correlates with single-cell diversity and prognosis in colorectal liver metastasis. *The Journal of experimental medicine*. **217**, e20191847 https://doi.org/10.1084/jem.20191847 (2020).
22. Pagès, F., et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet.* **391**, 2128–2139 https://doi.org/10.1016/S0140-6736(18)30789-X (2018).
23. Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation in *International Conference on Medical image computing and computer-assisted intervention* 234-241 (Springer, 2015).
24. Badrinarayanan, V., Kendall, A., & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2481-2495 https://doi.org/10.1109/TPAMI.2016.2644615 (2017).

25. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. https://arxiv.org/abs/1606.00915v2 (2016).
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). 4510-4520 https://arxiv.org/abs/1801.04381 (2018)
27. Falk, T., et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nature methods*, **16**, 67–70 https://doi.org/10.1038/s41592-018-0261-2 (2019).
28. Zeng, Z., Xie, W., Zhang, Y., Lu, Y. Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images. *IEEE Access* **7**; 21420-21427 doi 10.1109/ACCESS.2019.2896920 (2019)
29. De Brabandere, B., Neven, D., & Van Gool, L. "Semantic Instance Segmentation with a Discriminative Loss Function". https://arxiv.org/abs/1708.02551 (2017).
30. Hegde, P. S., & Chen, D. S. Top 10 Challenges in Cancer Immunotherapy. *Immunity*. **52,** 17–35 https://doi.org/10.1016/j.immuni.2019.12.011 (2020).
31. Bruni, D., Angell, H.K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat Rev Cancer.* https://doi.org/10.1038/s41568-020-0285-7 (2020).
32. Marliot, F., et al. Analytical validation of the Immunoscore and its associated prognostic value in patients with colon cancer. *Journal for immunotherapy of cancer.* **8**, e000272 https://doi.org/10.1136/jitc-2019-000272 (2020).
33. Quezada-Marín, J. I., et al. Gastrointestinal tissue-based molecular biomarkers: a practical categorisation based on the 2019 World Health Organization classification of epithelial digestive tumours. *Histopathology.* **77**, 340–350 https://doi.org/10.1111/his.14120 (2020).
34. Bera, K., Schalper K.A., Rimm D.L., Velcheti V., Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*. **16,** 703–715. https://doi.org/10.1038/s41571-019-0252-y (2019)
35. Shorten, C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *Journal of Big Data.* **6**, 60 https://link.springer.com/article/10.1186/s40537-019-0197-0 (2019).
36. Chen, L.C., Papandreou, G., Schroff, F., Adam, H. Rethinking atrous convolution for semantic image segmentation. https://arxiv.org/abs/1706.05587 (2017).
37. Kingma, D.P., Ba, J.A. A method for stochastic optimization. https://arxiv.org/abs/1412.6980 (2014).

AUTHOR CONTRIBUTIONS

PC: study design; pipeline development (slide annotation; data augmentation; CNN testing; data elaboration; metrics evaluation); manuscript writing;

NC, FM: study design; manuscript writing;

MD: case selection; study design;

CS, MDM: staining and annotation of slides;

MDS, LC: supervision pipeline development;

VS, GT, AM, LT, MR: critical review of manuscript

LDT: study design; supervision histological aspects (annotation; segmentation output); manuscript writing; study coordinator.

ADDITIONAL INFORMATION

The research leading to these results has received funding from Associazione Italiana per la Ricerca sul Cancro (AIRC) AIRC 5x1000 IG-21147 to AM and by Fondi 5x1000 anno 2019, progetto "ApPath: un dispositivo di patologia aumentata per valutare l'ambiente immune nel cancro colorettale metastatico" to LDT.

FIGURE LEGENDS.

Visual abstract of the study. The background. We recently proved that in human colorectal cancer, the presence of small and large tumor-associated macrophages (TAMs), is associated with different outcomes. **The challenge**. To translate this biological data into a robust clinical marker means to identify all TAMs in a single slide, which largely outnumber hundreds of cells, and then evaluate the area of each of them, a task unfeasible in the routine pathology workout. **Study design**. With the aim to develop a deep-learning pipeline to tackle this challenge, we selected, trained and tested three different approaches. **The solution**. The deep-learning pipeline based on the DeepLab-v3 architecture and semantic segmentation technique warrants the separation of TAMs from the background and the identification of single TAMs: this will easily allow the evaluation of their area.

Figure 1. TAM segmentation: semantic approach. The figure compares the histopathological image and the output of CNN model (UNet in these examples) using a semantic approach. In the original image TAM are stained in brown by CD163 immunostaining and the surrounding hepatocytes (the background) are counterstained in blue; (nuclei darker than cytoplasm); in CNN output TAM are returned in white and the background in black. The **semantic baseline analysis** generated an optimal separation between TAM and the background but was not able to separate adjacent TAMs, as highlighted by red arrows. Adopting the **semantic penalty** approach fixed part of this specific problem: the two TAM shown on the right are properly characterized by this solution.

Figure 2. TAM segmentation: instance approach. This figure compares the histopathological image and the outputs given by two models (UNet and DeepLab-v3) using instance segmentation. The output generated by Unet is characterized several instances (each characterized by a specific color) outnumbering true TAM. By contrast, Deep Lab-v3 model properly labelled each TAM as single instance, showing also a good performance to separate adjacent cells.

Figures

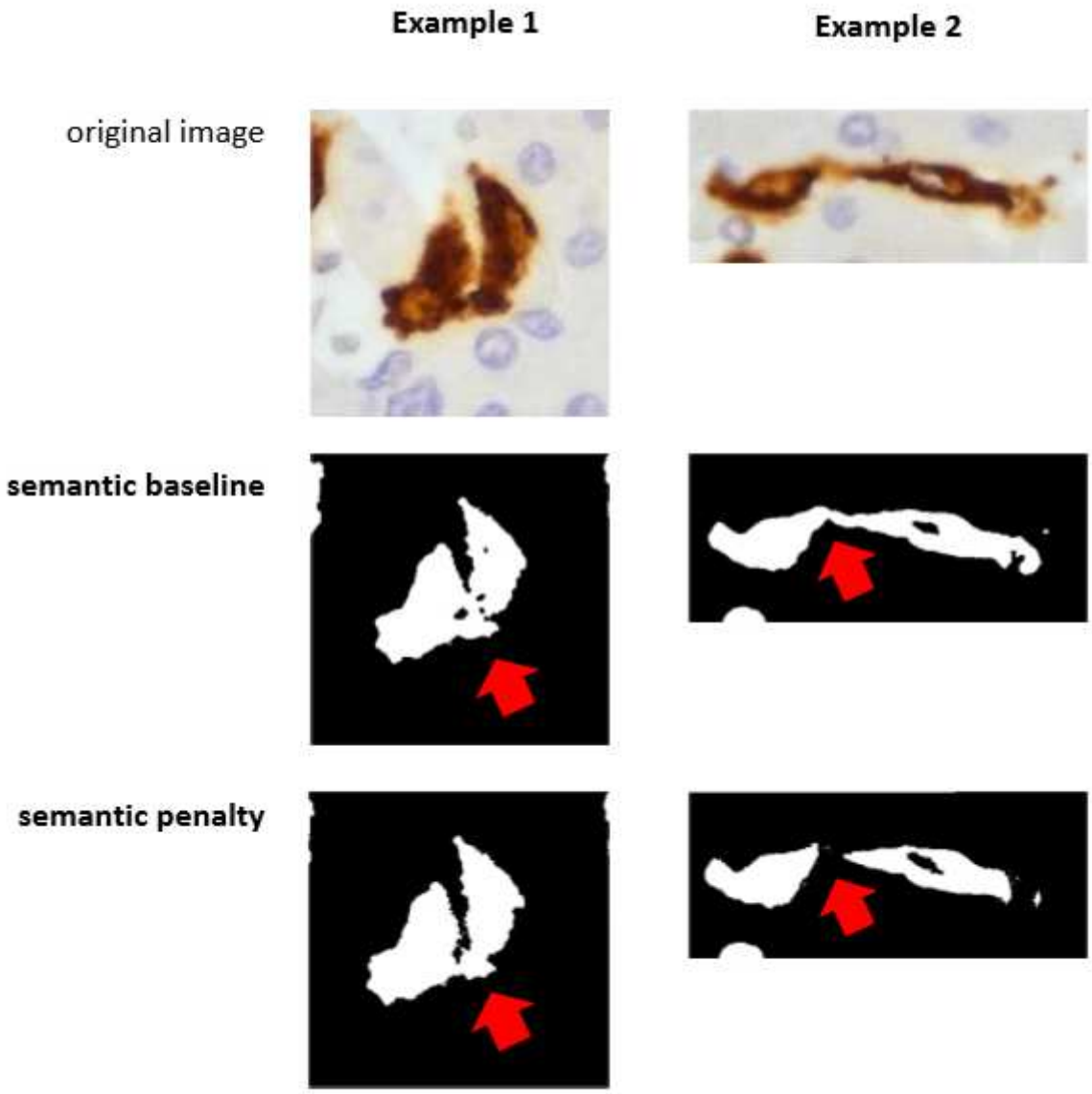


Figure 1

TAM segmentation: semantic approach. The figure compares the histopathological image and the output of CNN model (UNet in these examples) using a semantic approach. In the original image TAM are stained in brown by CD163 immunostaining and the surrounding hepatocytes (the background) are counterstained in blue; (nuclei darker than cytoplasm); in CNN output TAM are returned in white and the background in black. The semantic baseline analysis generated an optimal separation between TAM and the background but was not able to separate adjacent TAMs, as highlighted by red arrows. Adopting the semantic penalty approach fixed part of this specific problem: the two TAM shown on the right are properly characterized by this solution.

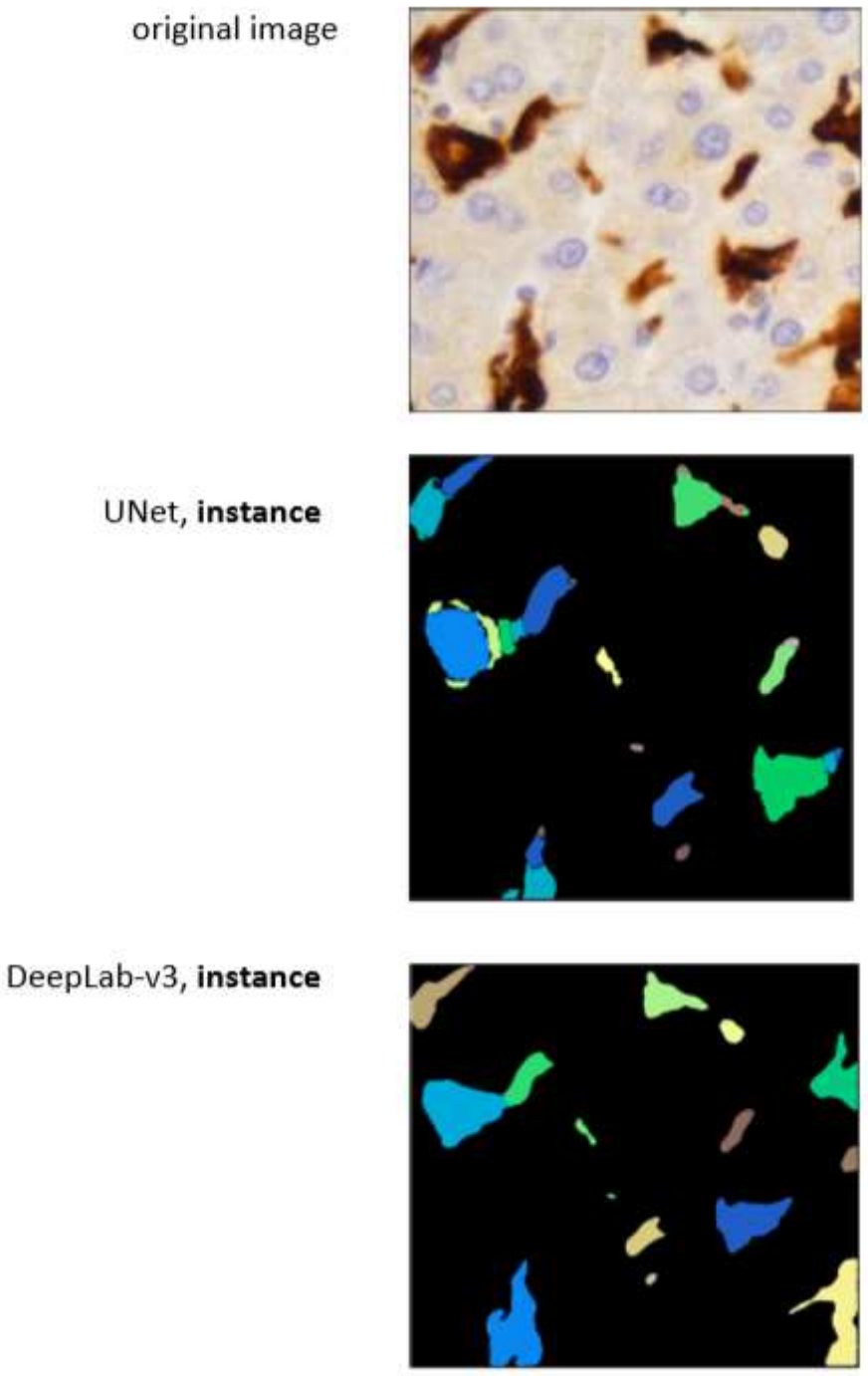


Figure 2

TAM segmentation: instance approach. This figure compares the histopathological image and the outputs given by two models (UNet and DeepLab-v3) using instance segmentation. The output generated by Unet is characterized several instances (each characterized by a specific color) outnumbering true TAM. By contrast, Deep Lab-v3 model properly labelled each TAM as single instance, showing also a good performance to separate adjacent cells.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Graphicalabstract.png