Biomedical Image Processing and Classification

Table 2. Comparison between the RENFAST algorithm and the current state of the art for blood vessel segmentation (pixel-based metrics).

Method	Subset	Comp. Time (s)	$Bal_{ACCURACY}$	Precision	Recall	$F1_{SCORE}$
Bevilacqua et al. [8]	TRAIN	2.58 ± 1.24	0.6845 ± 0.1467	0.8618 ± 0.1955	0.5115 ± 0.2196	0.5996 ± 0.1931
	TEST	2.64 ± 1.18	0.6487 ± 0.1494	0.7677 ± 0.2647	0.4944 ± 0.2241	0.5684 ± 0.2281
Two-class CNN [1]	TRAIN	**0.57 ± 0.11**	0.8821 ± 0.1116	0.9203 ± 0.0945	0.8026 ± 0.1630	0.8430 ± 0.1242
	TEST	**0.56 ± 0.09**	0.8116 ± 0.1305	0.9308 ± 0.1004	0.6923 ± 0.1743	0.7741 ± 0.1419
Three-class CNN [2]	TRAIN	0.74 ± 0.16	0.8744 ± 0.0861	**0.9888 ± 0.0337**	0.7706 ± 0.1199	0.8601 ± 0.0919
	TEST	0.71 ± 0.18	0.8220 ± 0.1075	**0.9800 ± 0.0800**	0.6666 ± 0.1837	0.7740 ± 0.1597
RENFAST algorithm	TRAIN	2.67 ± 0.41	**0.9443 ± 0.0821**	0.9185 ± 0.0634	**0.9151 ± 0.0950**	**0.9126 ± 0.0611**
	TEST	2.59 ± 0.53	**0.8936 ± 0.0969**	0.9269 ± 0.0845	**0.8185 ± 0.1344**	**0.8593 ± 0.0858**

[1] CNN with the same architecture shown in Figure 2 but trained on two classes (background vs. vessel). [2] Same deep network of the RENFAST algorithm but without post-processing (Section 2.4).

Table 3. Object-based metrics calculated on detected blood vessels for both the TRAIN and TEST sets.

Method	Subset	DSC	HD95 (μm)
Bevilacqua et al. [8]	TRAIN	0.7476 ± 0.1517	20.33 ± 21.67
	TEST	0.7668 ± 0.1381	22.31 ± 34.62
Two-class CNN [1]	TRAIN	0.7447 ± 0.2312	21.13 ± 30.59
	TEST	0.6879 ± 0.2417	26.68 ± 36.50
Three-class CNN [2]	TRAIN	0.7802 ± 0.1777	12.02 ± 22.45
	TEST	0.7483 ± 0.1790	9.35 ± 8.84
RENFAST algorithm	TRAIN	**0.8441 ± 0.1762**	**9.78 ± 10.51**
	TEST	**0.8358 ± 0.1391**	**6.41 ± 6.25**

[1] CNN with the same architecture shown in Figure 2 but trained on two classes (background vs. vessel). [2] Same deep network of the RENFAST algorithm but without post-processing (Section 2.4).

Regarding pixel-based metrics, our method achieved the best $Bal_{ACCURACY}$, recall, and $F1_{SCORE}$ for both the TRAIN and TEST sets. A large margin was achieved by RENFAST compared to the state-of-the-art techniques. Even more interesting, the post-processing adopted for blood vessel segmentation allowed a further increase in the overall performance of the single deep network (three-class CNN vs. RENFAST). The combination of the CNN probability map and cellular structure segmentation increased the DSC by up to 14.8% with respect to other methods. The accurate segmentation of blood vessel boundaries is also demonstrated by the lower HD95 value. Figure 7 shows a visual comparison between RENFAST and previously published works. Our approach managed to separate and correctly outline the boundaries of the blood vessels.

3.2. Fibrosis Segmentation

The same pixel-based metrics employed in the last section were calculated to evaluate the performance of RENFAST in fibrosis quantification (Table 4). To demonstrate the importance of the stain normalization as a preprocessing step, we also evaluated the performance of our algorithm without normalizing the images ("No norm.").

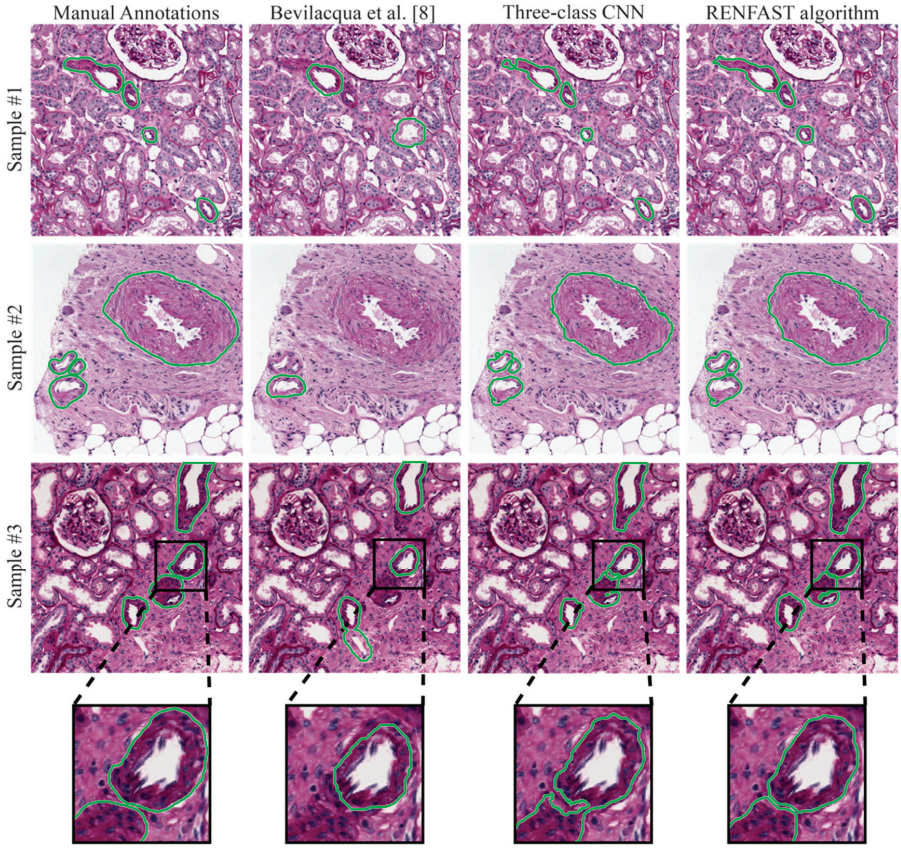


Figure 7. Blood vessel detection performed by state-of-the-art methods and the proposed algorithm. Two different samples are displayed in the first rows, while the last row shows a zoom of the segmentation near the blood vessel contour.

Table 4. Comparison between the proposed algorithm and the current state of the art for fibrosis segmentation (pixel-based metrics).

Method	Subset	Comp. Time (s)	$\text{Bal}_{\text{ACCURACY}}$	Precision	Recall	F1_{SCORE}
Tey et al. [10]	TRAIN	0.24 ± 0.04	0.8575 ± 0.0374	0.7538 ± 0.0780	0.8905 ± 0.0744	0.8147 ± 0.0515
	TEST	0.25 ± 0.07	0.8604 ± 0.0428	0.7512 ± 0.0736	0.9055 ± 0.0734	0.8166 ± 0.0492
Fu et al. [11]	TRAIN	**0.16 ± 0.06**	0.8988 ± 0.0660	0.8832 ± 0.1072	0.8940 ± 0.1469	0.8727 ± 0.0896
	TEST	**0.18 ± 0.09**	0.9159 ± 0.0491	0.8783 ± 0.1019	**0.9239 ± 0.1026**	0.8911 ± 0.0644
No norm. [1]	TRAIN	0.17 ± 0.07	0.9128 ± 0.0221	0.9025 ± 0.0482	0.8765 ± 0.0434	0.8900 ± 0.0240
	TEST	0.18 ± 0.11	0.9164 ± 0.0247	0.9157 ± 0.0304	0.8738 ± 0.0499	0.8944 ± 0.0277
RENFAST algorithm	TRAIN	0.27 ± 0.13	**0.9212 ± 0.0199**	**0.9064 ± 0.0355**	**0.8958 ± 0.0480**	**0.8973 ± 0.0275**
	TEST	0.29 ± 0.14	**0.9227 ± 0.0222**	**0.9184 ± 0.0313**	0.8891 ± 0.0482	**0.9010 ± 0.0246**

[1] RENFAST algorithm without the stain normalization as preprocessing.

As shown in Table 4, our strategy outperformed all the previously published methods. In addition, the stain normalization (Section 2.2) allowed a further increase in the overall performance of our method (No norm. vs. RENFAST algorithm). Finally, we evaluated the absolute errors (AEs) between the manual and automatic fibrosis quantification (Table 5). In both the TRAIN and TEST datasets, the RENFAST algorithm achieved the lowest average AEs (2.42% and 2.32%), with maximum AEs

of 11.17% and 7.81%, respectively. Specifically, the maximum AE obtained by our method was 3–5 times lower compared to state-of-the-art techniques [10,11]. Figure 8 shows some kidney fibrosis segmentation results.

Table 5. Minimum, average, and maximum absolute errors (AE_{MIN}, AE_{MEAN}, AE_{MAX}) between manual and automatic fibrosis quantification.

Method	Subset	AE_{MIN} (%)	AE_{MEAN} (%)	AE_{MAX} (%)
Tey et al. [10]	TRAIN	0.03	8.79	42.46
	TEST	0.59	8.73	38.41
Fu et al. [11]	TRAIN	0.01	7.81	38.62
	TEST	0.04	5.93	28.73
No norm. [1]	TRAIN	0.01	2.52	11.21
	TEST	0.05	2.50	8.29
RENFAST algorithm	TRAIN	**0.01**	**2.42**	**11.17**
	TEST	**0.01**	**2.32**	**7.81**

[1] RENFAST algorithm without the stain normalization as preprocessing.

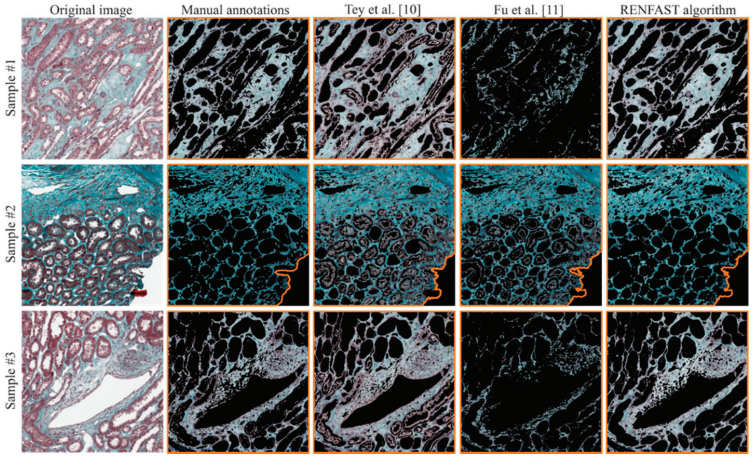


Figure 8. Visual performance comparison between previously published papers for fibrosis detection and the RENFAST algorithm. The fibrosis mask is superimposed on the original image, while the tissue contour is highlighted in orange.

3.3. Whole Slide Analysis

Since arteriosclerosis and fibrosis are generally assessed on whole slide images (WSIs), we extended our strategy to entire biopsies using a sliding window approach. To evaluate the degree of arterial sclerosis and fibrosis, an expert pathologist takes at least 20 min per patient, while the RENFAST algorithm is able to process the entire WSI in about 2 min. Figure 9 illustrates the results obtained using our algorithm on two different kidney biopsies stained with PAS (vessel detection) and TRIC (fibrosis segmentation). The introduction of an automatic algorithm within the clinical workflow can speed up the diagnostic process and provide more accurate data to assess kidney transplantability.

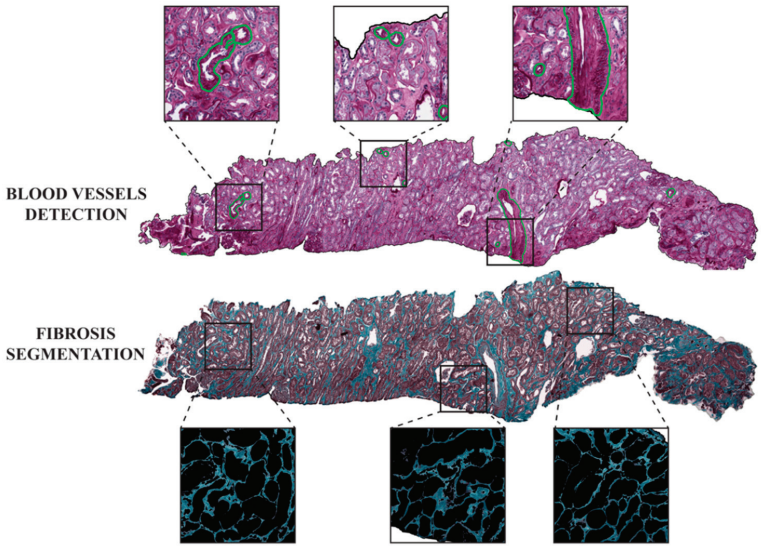


Figure 9. The result of RENFAST processing on a whole slide image (WSI). Blood vessels are shown in green in PAS stained WSIs. During the assessment of fibrosis, the connective tissue is segmented by removing all the tubular, vascular, and glomerular structures.

4. Discussion and Conclusions

Advances in transplant patient management are steadily increasing with improved clinical data and outcomes, requiring proportional development of the technical procedures routinely applied. However, the histopathological evaluation of preimplantation donor kidney biopsies has not varied, despite the increasing demand for pathology reports.

In this study, we present a fast and accurate method for the segmentation of kidney blood vessels and fibrosis in histological images. The detection of vascular structures and interstitial fibrosis is a real challenge due to the stain variability that affects the PAS and TRIC images, combined with high variation in the shape, size, and internal architecture of the renal structures. Thanks to the stain normalization step, our approach is capable of automatically detecting fibrotic areas and blood vessels in images with different staining intensity. The proposed algorithm was developed and tested on 350 PAS images for blood vessel segmentation and on 300 TRIC stained images for the detection of renal fibrosis. The results were compared with both manual annotations and previously published methods [8,10,11].

In blood vessel detection, the RENFAST algorithm achieved the best $\text{Bal}_{\text{ACCURACY}}$, recall, and F1_{SCORE} compared to other techniques. More importantly, our strategy obtained the best DSC and HD95 in the segmentation of vessel boundaries (Table 3). This is fundamental as accurate segmentation of the blood vessel borders is mandatory for the correct evaluation of vascular damage. This high performance is mainly due to the combination of CNN segmentation with ad hoc post-processing specifically designed to detect the contour of each blood vessel. By segmenting lumen regions and cell nuclei, the RENFAST algorithm manages to delete almost all the false-positive shapes detected by the CNN. Our strategy is also capable of segmenting small blood vessels and correctly separating touching structures (Figure 7).

On TRIC stained images, the RENFAST algorithm allows us to quantify the interstitial fibrosis. The proposed approach showed high accuracy in segmenting fibrotic tissue and outperformed all the previously published methods (Table 4). Compared with the current state-of-the-art techniques, our method obtained the lowest absolute error (around 2.4%) in the estimation of fibrosis percentage.

In the TEST set, the maximum absolute error of the algorithm was only 7.81%, about 4 times lower with respect to the compared methods. The combination of color normalization and adaptive stain separation allows us to accurately quantify the extent of the fibrotic area.

Although the proposed strategy is fast and robust, it still has some limitations. First of all, the histological images must be acquired at 10× or higher magnification. Using a lower resolution (5× or below), the deep network cannot accurately segment the blood vessels, and cell nucleus segmentation may fail due to the poor quality of the image. Another limitation refers to the WSI application. Nowadays, pathologists evaluate only arteriolar narrowing and interstitial fibrosis in the renal cortex, excluding all structures of the medulla from the evaluation. Our algorithm does not yet include a pipeline for the recognition of the medullary tissue from the cortical tissue on kidney biopsies. However, its potential in assessing vessel and parenchyma injury represents an efficient tool to increase accuracy, reproducibility, and velocity in an increasingly urgent medical setting.

In this study, we presented a simple yet effective pipeline for blood vessel and fibrosis segmentation in kidney histological images. Our research group is currently working on the extension of the RENFAST algorithm to automatically detect the cortical tissue on WSIs and assign a vascular score according to [5]. In the future, we will integrate the assessment of glomerulosclerosis and tubular atrophy within the RENFAST algorithm in order to create the first automated Karpinski scoring system.

Author Contributions: Conceptualization, L.M. and F.M.; methodology, M.S.; software, M.S. and A.M.; validation, A.M. and K.M.M.; resources, A.G. and A.B.; data curation, A.G. and L.M.; writing—original draft preparation, M.S.; writing—review and editing, A.M. and K.M.M.; supervision, M.P. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to acknowledge all the laboratory technicians of the Division of Pathology (Department of Oncology, Turin, Italy) for their help in digitizing histological slides.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

During the inference phase, the CNN's probability map could suffer from a lack of information near the edges of the image. To overcome this problem, an *Extended image* is synthesized by padding the original image with mirror reflections of 256 × 256 pixels along each direction. As shown in Figure A1, the result of this operation is an RGB image of 1024 × 1024 pixels. A sliding window operator with a size of 512 × 512 is then passed over the extended image with an overlap of 256 pixels between consecutive windows. The deep network is applied to each 512 × 512 window, and only the center of each prediction is kept for the creation of the initial softmax. This operation yields a heat map of size 768 × 768 which is further center cropped to obtain the final softmax with the same size as the input image. The final softmax can be considered as an RGB image, where the red layer contains the probability for each pixel of belonging to the "blood vessel" class, while the green layer represents the probability for each pixel of belonging to the "blood vessel boundaries" class.

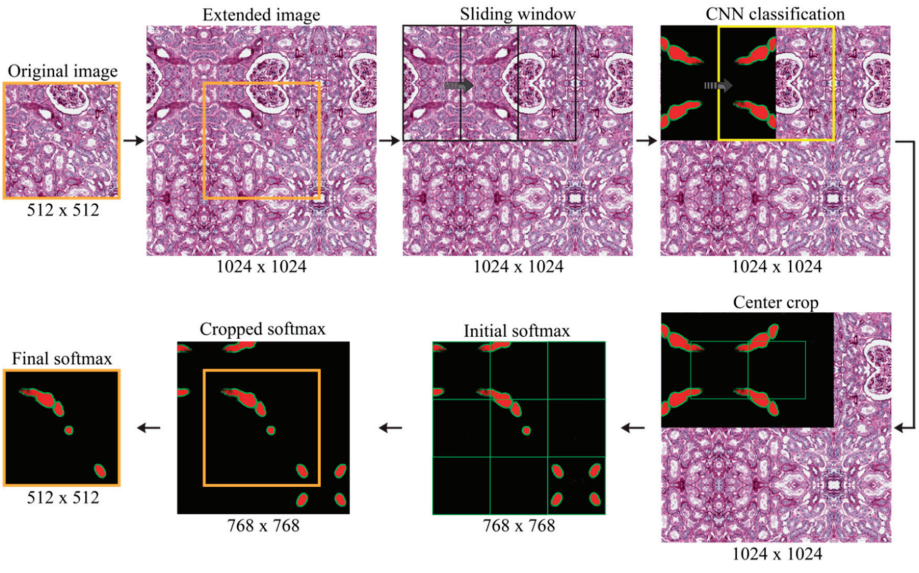


Figure A1. Procedure for the creation of the final CNN softmax. The original image is mirrored around the boundaries to obtain the extended image. Then, a sliding window approach is employed to classify each patch, and only the center of each prediction is kept to build the final softmax.

Appendix B

The semi-automatic pipeline used to generate the manual annotation of fibrotic areas was developed in Fiji [20]. Fiji is a Java-based software product with several plugins that facilitate medical image analysis. The proposed pipeline consists of seven steps: (i) image loading; (ii) manual definition of a ROI (region of interest) for each of the three colors (white, green, red); (iii) RGB color averaging of each ROI to obtain the three stain vectors; (iv) color deconvolution using the stain vectors previously found; (v) manual thresholding on the green channel; (vi) small particle removal; and (vii) complementation of the binary mask.

References

1. Salmon, G.; Salmon, E. Recent Innovations in Kidney Transplants. *Nurs. Clin. N. Am.* **2018**, *53*, 521–529. [CrossRef] [PubMed]
2. Metzger, R.A.; Delmonico, F.L.; Feng, S.; Port, F.K.; Wynn, J.J.; Merion, R.M. Expanded criteria donors for kidney transplantation. *Am. J. Transplant.* **2003**, *3*, 114–125. [CrossRef] [PubMed]
3. Heilman, R.L.; Mathur, A.; Smith, M.L.; Kaplan, B.; Reddy, K.S. Increasing the use of kidneys from unconventional and high-risk deceased donors. *Am. J. Transplant.* **2016**, *16*, 3086–3092. [CrossRef]
4. Altini, N.; Cascarano, G.D.; Brunetti, A.; Marino, F.; Rocchetti, M.T.; Matino, S.; Venere, U.; Rossini, M.; Pesce, F.; Gesualdo, L. Semantic Segmentation Framework for Glomeruli Detection and Classification in Kidney Histological Sections. *Electronics* **2020**, *9*, 503. [CrossRef]
5. Karpinski, J.; Lajoie, G.; Cattran, D.; Fenton, S.; Zaltzman, J.; Cardella, C.; Cole, E. Outcome of kidney transplantation from high-risk donors is determined by both structure and function. *Transplantation* **1999**, *67*, 1162–1167. [CrossRef] [PubMed]
6. Carta, P.; Zanazzi, M.; Caroti, L.; Buti, E.; Mjeshtri, A.; Di Maria, L.; Raspollini, M.R.; Minetti, E.E. Impact of the pre-transplant histological score on 3-year graft outcomes of kidneys from marginal donors: A single-centre study. *Nephrol. Dial. Transplant.* **2013**, *28*, 2637–2644. [CrossRef]
7. Furness, P.N.; Taub, N.; Project, C. of E.R.T.P.A.P. (CERTPAP) International variation in the interpretation of renal transplant biopsies: Report of the CERTPAP Project. *Kidney Int.* **2001**, *60*, 1998–2012. [CrossRef]

8. Bevilacqua, V.; Pietroleonardo, N.; Triggiani, V.; Brunetti, A.; Di Palma, A.M.; Rossini, M.; Gesualdo, L. An innovative neural network framework to classify blood vessels and tubules based on Haralick features evaluated in histological images of kidney biopsy. *Neurocomputing* **2017**, *228*, 143–153. [CrossRef]
9. He, D.-C.; Wang, L. Texture features based on texture spectrum. *Pattern Recognit.* **1991**, *24*, 391–399. [CrossRef]
10. Tey, W.K.; Kuang, Y.C.; Ooi, M.P.-L.; Khoo, J.J. Automated quantification of renal interstitial fibrosis for computer-aided diagnosis: A comprehensive tissue structure segmentation method. *Comput. Methods Programs Biomed.* **2018**, *155*, 109–120. [CrossRef] [PubMed]
11. Fu, X.; Liu, T.; Xiong, Z.; Smaill, B.H.; Stiles, M.K.; Zhao, J. Segmentation of histological images and fibrosis identification with a convolutional neural network. *Comput. Biol. Med.* **2018**, *98*, 147–158. [CrossRef]
12. Monaco, J.; Hipp, J.; Lucas, D.; Smith, S.; Balis, U.; Madabhushi, A. Image segmentation with implicit color standardization using spatially constrained expectation maximization: Detection of nuclei. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Nice, France, 1–5 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 365–372.
13. Peter, L.; Mateus, D.; Chatelain, P.; Schworm, N.; Stangl, S.; Multhoff, G.; Navab, N. Leveraging random forests for interactive exploration of large histological images. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Boston, MA, USA, 14–18 September 2014; Springer: Cham, Switzerland, 2014; pp. 1–8.
14. Ciompi, F.; Geessink, O.; Bejnordi, B.E.; De Souza, G.S.; Baidoshvili, A.; Litjens, G.; Van Ginneken, B.; Nagtegaal, I.; Van Der Laak, J. The importance of stain normalization in colorectal tissue classification with convolutional networks. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 160–163.
15. Salvi, M.; Michielli, N.; Molinari, F. Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. *Comput. Methods Programs Biomed.* **2020**, *193*, 105506. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
18. Salvi, M.; Molinari, F. Multi-tissue and multi-scale approach for nuclei segmentation in H&E stained images. *Biomed. Eng. Online* **2018**, *17*. [CrossRef]
19. Salvi, M.; Molinaro, L.; Metovic, J.; Patrono, D.; Romagnoli, R.; Papotti, M.; Molinari, F. Fully automated quantitative assessment of hepatic steatosis in liver transplants. *Comput. Biol. Med.* **2020**, *123*, 103836. [CrossRef]
20. Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B. Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **2012**, *9*, 676–682. [CrossRef]

electronics

MDPI

Article

Deep Learning with Limited Data: Organ Segmentation Performance by U-Net

Michelle Bardis [1,*], Roozbeh Houshyar [1], Chanon Chantaduly [1], Alexander Ushinsky [2], Justin Glavis-Bloom [1], Madeleine Shaver [1], Daniel Chow [1], Edward Uchio [3] and Peter Chang [1]

1 Department of Radiological Sciences, University of California, Irvine, CA 92617, USA; rhoushya@hs.uci.edu (R.H.); cchantad@hs.uci.edu (C.C.); jglavisb@hs.uci.edu (J.G.-B.); mshaver@hs.uci.edu (M.S.); chowd3@hs.uci.edu (D.C.); changp6@hs.uci.edu (P.C.)

2 Mallinckrodt Institute of Radiology, Washington University Saint Louis, St. Louis, MO 63110, USA; aushinsky@wustl.edu

3 Department of Urology, University of California, Orange, CA 92868, USA; euchio@hs.uci.edu

* Correspondence: mbardis@hs.uci.edu

Received: 19 June 2020; Accepted: 23 July 2020; Published: 26 July 2020

Abstract: (1) Background: The effectiveness of deep learning artificial intelligence depends on data availability, often requiring large volumes of data to effectively train an algorithm. However, few studies have explored the minimum number of images needed for optimal algorithmic performance. (2) Methods: This institutional review board (IRB)-approved retrospective review included patients who received prostate magnetic resonance imaging (MRI) between September 2014 and August 2018 and a magnetic resonance imaging (MRI) fusion transrectal biopsy. T2-weighted images were manually segmented by a board-certified abdominal radiologist. Segmented images were trained on a deep learning network with the following case numbers: 8, 16, 24, 32, 40, 80, 120, 160, 200, 240, 280, and 320. (3) Results: Our deep learning network's performance was assessed with a Dice score, which measures overlap between the radiologist's segmentations and deep learning-generated segmentations and ranges from 0 (no overlap) to 1 (perfect overlap). Our algorithm's Dice score started at 0.424 with 8 cases and improved to 0.858 with 160 cases. After 160 cases, the Dice increased to 0.867 with 320 cases. (4) Conclusions: Our deep learning network for prostate segmentation produced the highest overall Dice score with 320 training cases. Performance improved notably from training sizes of 8 to 120, then plateaued with minimal improvement at training case size above 160. Other studies utilizing comparable network architectures may have similar plateaus, suggesting suitable results may be obtainable with small datasets.

Keywords: training size; deep learning; convolutional neural network; U-Net; segmentation; artificial intelligence

1. Introduction

Deep learning through convolutional neural networks (CNNs), a subset of artificial intelligence, has demonstrated many strengths for image analysis [1]. For example, CNN approaches represent all recent winning entries within the annual ImageNet Classification challenge, consisting of over one million photographs in 1000 object categories with a 3.6% classification error rate to date [2,3]. In addition, medical applications have demonstrated potential to improve triage with intracranial hemorrhage detection [4] and glioma genetic mutation classification [5]. However, a CNN's performance depends on its ability to learn from the input data itself, and a CNN requires both (1) high-quality and (2) large datasets to solve problems effectively [6,7]. By determining the relationship between dataset size and CNN accuracy, investigators could potentially calculate when a CNN has been effectively trained.

Training data scarcity and quality are generally not considered challenges for non-biomedical applications where data is widely available. For example, Facebook collects more than 50 TB of video per day and Google processes 200,000 TB per day [8,9]. By contrast, biomedical datasets tend to be heterogenous, difficult to annotate, and relatively scarce [10,11]. In two recent breast imaging studies that used artificial intelligence (AI), the dataset sizes for breast lesion detection and breast cancer recurrence were 320 and 92 patients, respectively [12,13]. Medical studies often lack a combination of publicly available data and high-quality labels [1,14]. Recognition of rare diseases proves especially challenging for medical imaging neural networks, as imaging data for these diseases are often very limited [14]. Additionally, annotation of clinical data is a time consuming and potentially expensive process. Consequently, most medical imaging CNNs face a scarcity of data and calculating an optimal dataset size is infeasible [14].

Since most medical imaging studies are constrained by small datasets, few studies have examined the relationship between the number of cases and CNN performance. A study by Cho et al. [15] compared the number of cases versus performance for a CNN that classified axial computerized tomography scans (CTs) into different anatomic regions: brain, neck, shoulder, chest, abdomen, and pelvis. Another study by Lakhani et al. [16] also observed the performance difference with four different case sizes for CNNs that identified the presence or absence of an endotracheal tube on chest radiographs. Although these two studies showed better accuracy with more cases, the CNNs utilized in the studies completed image classification tasks that make a binary decision after examining the image in its entirety. The relationship between number of cases and segmentation performance within an image has not been rigorously explored.

The purpose of this study is to identify the ideal training size for prostate organ segmentation by analyzing the relationship between the number of MRI cases utilized and consequent CNN performance for imaging analysis. We implemented a type of CNN called a U-Net [17], which was specifically created for medical imaging assessment tasks typically lacking large datasets. U-Net is widely used in medical imaging artificial intelligence (AI) research. We hypothesize a plateau in performance because organ segmentation is a suitable and straightforward task for a U-Net.

2. Materials and Methods

2.1. Patient Selection

This retrospective study was granted a waiver of informed consent by the institutional review board (IRB) at the University of California, Irvine (UCI) for use of human subject data in a research study. An institutional prostate cancer database was searched to identify patients who had both a (1) prostate multiparametric magnetic resonance imaging (mpMRI) and a (2) magnetic resonance imaging/transrectal ultrasound (MRI/TRUS) fusion biopsy between September 2014 and August 2018. The inclusion criteria for this study included patients who had an mpMRI with subsequent 12-core Artemis 3D TRUS (Eigen, Grass Valley, CA, USA) and MRI/TRUS fusion biopsy using Artemis and ProFuse software (Eigen, Grass Valley, CA, USA) at the University of California, Irvine. An MRI/TRUS fusion biopsy was included as criteria because the prostate organ ground truth was segmented for these patients.

2.2. Image Acquisition

The mpMRI images were acquired on a Siemens Magnetom Trio 3-Tesla MRI scanner (Siemens AG, Munich, Germany) and a Phillips Ingenia 3-Tesla MRI scanner (Phillips Healthcare, Amsterdam, Netherlands) at the University of California, Irvine. The image acquisitions were completed in adherence to the prostate imaging reporting and data system (PI-RADS) v2 protocol without endorectal coil (Table 1).

Table 1. Magnetic resonance imaging (MRI) acquisition parameters.

Parameter	Measure
Field Strength (B0)	3 Tesla
Acquisition Technique	Turbo spin echo/echo planar
Echo Train Length	25
Time Repetition	7300 milliseconds
Time Echo	108 milliseconds
Flip Angle	150 degrees
Field of View	200 × 200 voxels
Matrix Size	256 × 205 pixels
Slice Thickness	3 mm
Slice Spacing	3 mm
Coil	Body

2.3. Ground Truth Segmentation

Ten radiologists manually segmented the prostate organ on axial T2-weighted (T2W) images with Profuse software (Eigen, Grass Valley, CA, USA). A board-certified abdominal radiologist with over 10 years of experience (R.H.) was the most experienced radiologist who approved each case. When other radiologists' segmentations differed from his expertise, he refined and updated those segmentations to establish the final ground truth. The mpMRI data and prostate organ segmentation data were transferred to a proprietary research database. From the database, the T2W axial images and organ segmentations were accessed and revised on an in-house image segmenting tool. The in-house tool enabled any segmentation corrections to be completed quickly. This tool integrated with the neural network training software and could be accessed with a web browser. Any segmentation updates were thus seamlessly updated into the neural network implementation.

2.4. Image Preprocessing

All axial images were resized to 256 × 256 voxels for neural network training. The axial slices were set to have a distance of 3 mm between each other. The standard deviation and mean values of each image were calculated when retrieved from the database. The image signal intensity was then normalized and applied voxelwise to each image. From all the available mpMRI sequences, only the T2W images were used for training and validation.

2.5. Convolutional Neural Network

The CNN used for this study was a custom modified U-Net. The algorithm's base architecture was derived from a standard U-Net, which is a fully convolutional contracting and expanding architecture [17]. The customized U-Net has a symmetric architecture and uses the same number of layers during subsampling and upsampling. U-Net also employs skip connections that allow the CNN to combine features for the image contraction and expansion pathways. These skip connections enabled the U-Net to use spatial information that could potentially be lost after the image is further downsampled in the contraction pathway. The entire image was trained during a single forward pass and the U-Net classified each image per pixel.

Our customized U-Net was extended to incorporate three dimensions during training and then produce outputs in two dimensions (Figure 1). Five layers were chosen empirically. In each layer, the image was processed by batch normalization, convolution, rectified linear unit (ReLU) activation, and downsampling with strided convolutions by a factor of 2. The 5 layers used 4, 8, 16, 32, and 32 filters per convolution. The image was downsampled until it became a 1 × 1 × 1 matrix before it underwent expansion. During the expansion pathway, the image was upsampled and a skip connection allowed the upsampled image to combine spatial information from the contraction pathway.

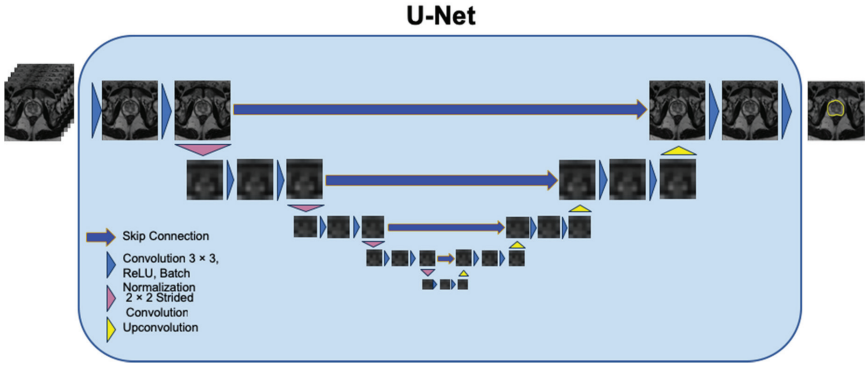


Figure 1. All neural network runs were completed on a U-Net with 5 layers. The number of channels used were 4, 8, 16, 32, and 32 for the 5 layers.

2.6. Algorithm Training

The Adam optimization algorithm was employed to update the network weights. The Adam algorithm used classical stochastic gradient descent during training [18]. The learning rate was set to 1×10^{-3}, while the exponential decay rates, β_1 and β_2, were set to 0.9 and 0.999, respectively. The batch size was set to 32. The U-Net was trained over a range of iterations: 12,000 to 96,000. The hyperparameters and network structure were kept constant across all 12 runs.

The CNN was written with TensorFlow r1.9 library (Apache 2.0 license) and Python 3.5. The neural network was trained on a graphics processing unit (GPU) workstation which employed four GeForce GTX 1080 Ti cards (11 GB, Pascal microarchitecture; NVIDIA, Santa Clara, CA, USA).

2.7. Statistical Methods

The U-Net performance was measured by examining the Dice score. X and Y are both spatial target regions and their overlap is defined by the Dice score:

$$\text{Dice} = \frac{2\,|\,X \cap Y\,|}{|X| + |Y|}. \tag{1}$$

The Dice score quantifies the spatial overlap between the manually segmented and neural network-derived segmentations (Appendix A, Figure A1). A Dice score ranges from 0 (no overlap) to 1 (perfect overlap). A Dice score is the most widely used metric for evaluating segmentation performance for a neural network [19]. To estimate the stability of the neural network during training, the variance of the training Dice score was calculated.

The total number of cases available for training and validation was 400 MRIs. Our U-Net was implemented for 12 runs and trained on the following number of cases: 8, 16, 24, 32, 40, 80, 120, 160, 200, 240, 280, and 320 cases. For each of the 12 runs, the cases were randomly partitioned as either training or validation and the entire set of 400 cases were used. The Dice score was calculated for every validation case in every run. From validation cases in every run, the mean and standard deviation of the Dice scores were computed. For example, the CNN in Run 1 was trained on 8 cases. After the CNN was done training, validation on 392 cases that produced 392 different Dice scores was completed. The mean and standard deviation for these 392 Dice scores were 0.424 and 0.206, respectively. Training size, validation size, mean Dice score, and standard deviation of Dice score are listed for Runs 1 through 12 in Table 2.

Table 2. Mean Dice score and standard deviation of Dice score for 12 training sizes.

Run	Training Size (Cases)	Validation Size (Cases)	Mean Dice Score	Standard Deviation of Dice Score
1	8	392	0.424	0.206
2	16	384	0.653	0.160
3	24	376	0.716	0.145
4	32	368	0.724	0.150
5	40	360	0.747	0.147
6	80	320	0.819	0.099
7	120	280	0.793	0.113
8	160	240	0.858	0.068
9	200	200	0.840	0.111
10	240	160	0.855	0.076
11	280	120	0.857	0.082
12	320	80	0.867	0.090

After calculating the mean Dice score for 12 different runs, the SciPy [20] library in Python was used to complete curve fitting to these 12 data points with three nonlinear functions (Equations (2)–(4)). Multiple functions were used to optimize the regression and the best three functions that approximate the data were shown (Equations (2)–(4)). These three functions were selected from the SciPy library because they most effectively modeled the dataset that increased quickly from training sizes 8 to 32 and then gradually from training sizes 200 to 320. For all three functions, a, b, and c were constants, y was the Dice score, and x was the training size (Figure 2). The first function was logarithmic, with the formula:

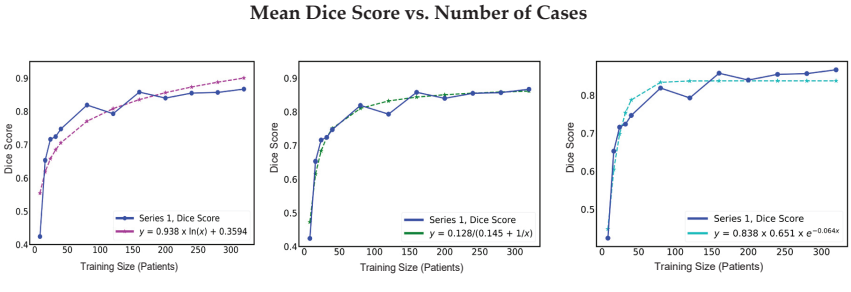
$$y = a \times \ln(x) + b. \tag{2}$$

The second function was asymptotic and used the formula:

$$y = \frac{a}{b + \frac{1}{x}}. \tag{3}$$

The third function was exponential, with the formula:

$$y = 1 - a \times e^{-b \times x} + c \tag{4}$$



(**a**) (**b**) (**c**)

Figure 2. The mean Dice score at 12 different training sizes was approximated with several curve functions. (**a**) The first function was logarithmic with the formula $y = a \times \ln(x) + b$. a was 0.938 and b was 0.3594. The mean squared error was 2.55×10^{-3}. (**b**) The second function was asymptotic and used the formula $y = \frac{a}{b+\frac{1}{x}}$. a was 0.128 and b was 0.145. The mean squared error was 5.70×10^{-4}. (**c**) The third function was exponential with the formula $y = 1 - a \times e^{-b \times x} + c$. a was 0.651, b was 0.064, and c was −0.162. The mean squared error was 8.30×10^{-4}. The second function (**b**) provided the best approximation because it had the lowest mean squared error.

For each approximation, the mean squared error was calculated with the following formula:

$$\text{Mean Squared Error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{y_i})^2 \tag{5}$$

where n was 12, y was the Dice score, and $\widetilde{y_i}$ was the estimated Dice score produced by the function.

3. Results

3.1. Prostate Segmentation

A total of 400 cases (10,400 axial images) from 374 patients were used during training and validation in our study. The average patient age was 65 years (range 41 to 96 years). The average prostate volume was 59 cm^3 (range 2 cm^3 to 353 cm^3). The relationship between number of cases used for training and algorithm performance is shown in Figure 3. The Dice score improved most when the case number changed from 8 to 16 (Table 2). In addition, the Dice score also started to plateau at a training size of 160 cases. The Dice score was 0.858 at 160 cases and 0.867 at 320 cases. To show progression of the Dice score, a single axial image from one case was selected to show the benefits of increasing the number of cases (Figure 4). On this one axial slice, the Dice score progressed from 0 to 0.98 as the training size grew from 8 to 320 cases.

Three nonlinear functions from the SciPy library were used to best fit the mean Dice score performance across the 12 runs. For the first function (2), a was 0.938, b was 0.3594, and the mean squared error was 2.55×10^{-3}. For the second function (3), a was 0.128, b was 0.145, and the mean squared error was 5.70×10^{-4}. For the third function (4), a was 0.651, b was 0.064, c was −0.162, and the mean squared error was 8.30×10^{-4}. The best curve fitting was completed by the second function and produced the lowest mean squared error.

Figure 3. Dice score improved the most between 8 cases and 16 cases (0.424 to 0.653). The Dice score started to plateau after 160 cases which had a performance of 0.858. The Dice score only improved by 0.09 from 160 cases to 320 cases. The Dice score was plotted with error bars that show the standard deviation above and below that run's mean Dice score. The standard deviation was lowest at 0.076 with 240 cases and highest at 0.206 with 8 cases.

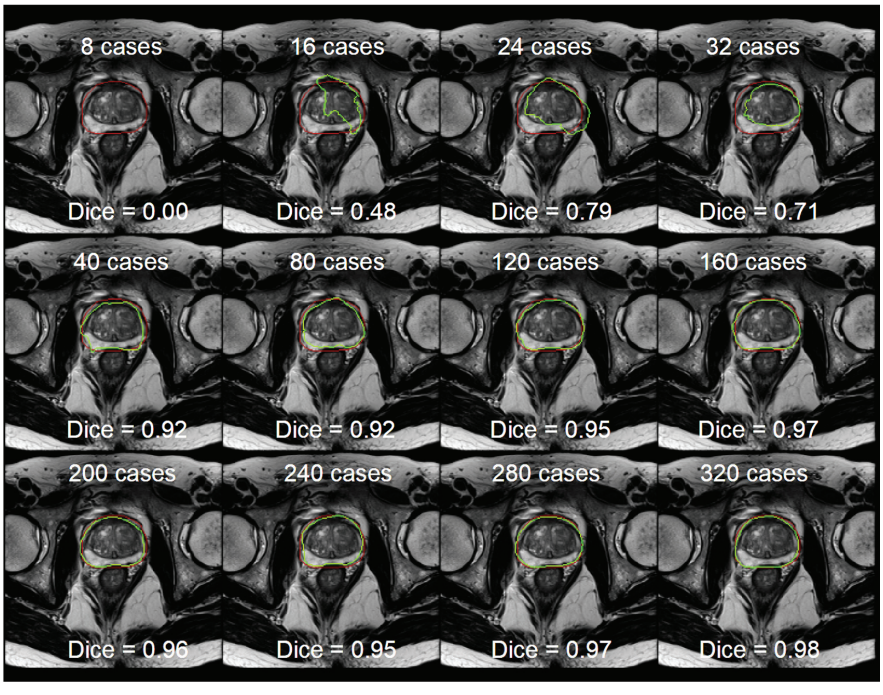


Figure 4. The performance of the U-Net was plotted for one axial slice on a single case across the different training sizes. The red line is the ground truth and the green line is the U-Net. The Dice score for one axial slice is shown in each square. The Dice score started to stabilize once the neural network trained with 160 cases.

3.2. Convolutional Neural Network Details/Statistics

The stability of the U-Net in training was evaluated (Figure 5). During training, the neural network runs that used training sizes between 8 and 40 did not converge quickly. By contrast, the neural network runs that used training sizes between 80 and 320 did converge quickly. The highest variance was 0.046 for the run that used 40 cases and the lowest variance was 0.003 for that run that used 200 cases. The training process required approximately 7 h of training time for each run. During inference, the U-Net took an average of 0.24 s per case on one GPU to complete inference.

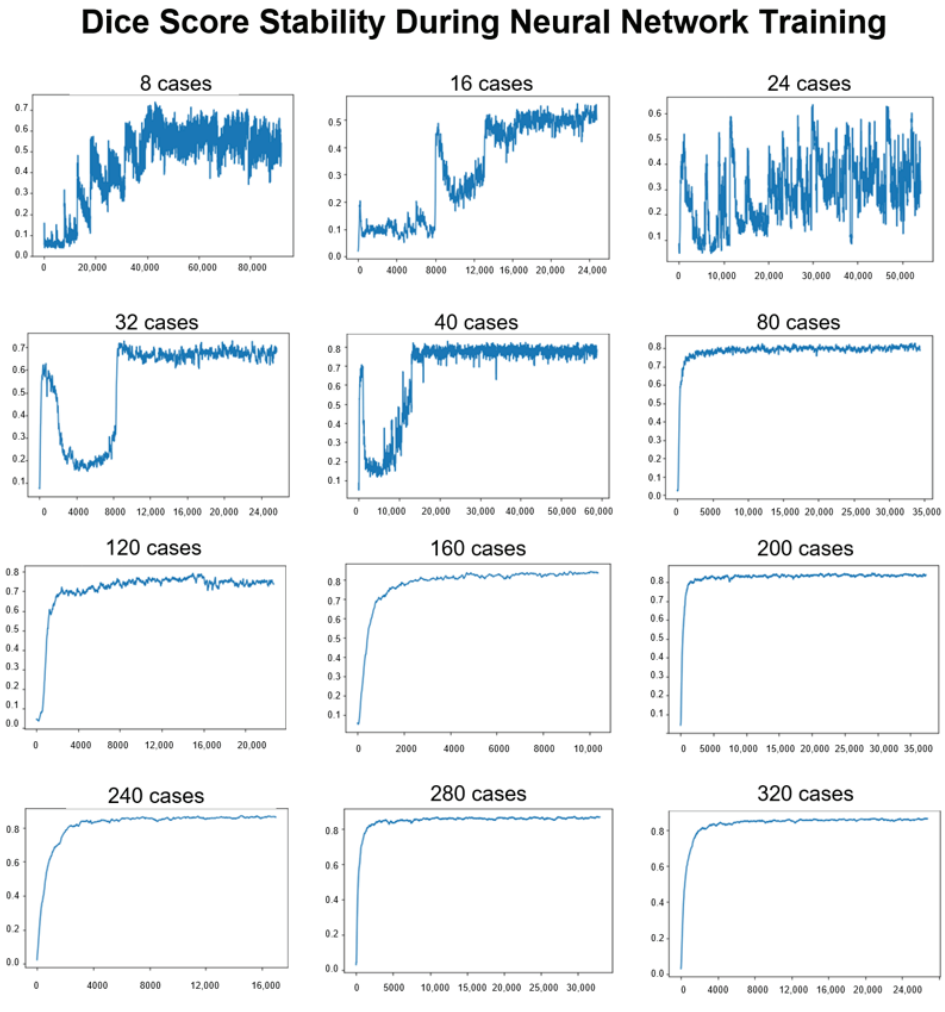


Figure 5. The number of iterations is plotted on the *x*-axis and the Dice score during training is plotted on the *y*-axis. The mean Dice score was plotted during training for the 12 different dataset sizes. The Dice score exhibited instability when training on case sizes of 8, 16, 24, 32, and 40. The Dice score stabilized more easily on case sizes of 80, 120, 160, 200, 240, 280, and 320. The Dice score variance was calculated during training; the run with 40 cases had the highest variance of 0.046 and the run with 200 cases had the lowest variance of 0.003.

4. Discussion

The purpose of this study was to explore the relationship between training size and CNN performance for prostate organ segmentation. As expected, the CNN performance plateaued with more data after 160 cases, providing a minimal increase in the Dice score. The Dice score was 0.858 at 160 cases and improved to 0.867 at 320 cases. These results confirm our hypothesis that providing more data after a certain size would only provide marginal benefits. The Dice score performance was best modeled with an asymptotic function (Equation (3)) that will converge as the number of cases increases. By using this asymptotic function (Equation (3)) for prediction, the Dice score would reach 0.871 with 500 cases and 0.877 with 1000 cases. The results also demonstrated that the selection of

U-Net as the CNN was apt due to effective prostate segmentation. U-Net's design that classifies each voxel after contraction and expansion are completed to extract unique features make it an apt network for medical imaging analysis. Since manual prostate segmentation is a tedious task [21] and took between 3 and 7 min per case for our radiologists, it is beneficial to know that more cases will not automatically translate into superior results.

Our study is unique because of its dataset size, which enabled us to find an optimal number of cases for training. In ten previous studies that also completed prostate segmentation, the dataset sizes ranged from 21 to 163 cases [22–31]. Three of these studies by Zhu et al. [28], Zhu et al. [27], and Clark et al. [26] were most comparable to our study because they also used a U-Net for their CNN. These three studies obtained Dice scores of 0.89, 0.93, and 0.89 with dataset sizes of 134, 163, and 81 cases, respectively. Although these studies did not compare training with multiple dataset sizes, their results support our findings that U-Net can achieve accurate results for prostate segmentation with a limited dataset.

Along with prostate segmentation, U-Net has demonstrated that it can segment other organs with small dataset sizes. The kidneys were accurately segmented by a U-Net in a study by Jackson et al. [32] with 89 cases. Jackson's study achieved Dice scores of 0.91 and 0.86 for the left and right kidneys, respectively [32]. Multiple U-Nets were combined together to segment multiple organs simultaneously on thorax computed tomography (CT) images in a study by Dong et al. [33]. In Dong's study, the network trained with 40 cases to obtain Dice scores of 0.97, 0.97, 0.90, and 0.87 for the left lung, right lung, spinal cord, and heart, respectively [33]. These studies demonstrate that a U-Net is a well-suited CNN for organ segmentation because of its ability to provide accurate results on small datasets. If these studies were to increase their number of cases, their Dice scores would probably improve and eventually plateau as well.

Several limitations should be considered in our study. All training data were gathered from one academic institution and two manufacturers' MRI scanners. All acquisitions were performed at 3 tesla (3T) MRI field strength and without endorectal coil. Although our CNN works well on our dataset, its ability to generalize with more prostate MRIs outside of our institution could be tested with studies from other institutions. Further work should explore the minimum amount of data for other tasks that build upon prostate organ segmentation. Different dataset sizes could be used to train networks that identify different prostate zones [34] and detect prostate lesions [35]. Along with the prostate, the training dataset size could be varied for other abdominal organs such as the kidney. These studies would serve as useful reference points for future studies that seek to optimize their neural networks. Additional work in this dataset should progress beyond prostate segmentation and detect prostate lesions. Lesion identification is a much more challenging task for AI and data augmentation with a generative adversarial network (GAN) [36] could be very useful since this technical problem lacks sufficient training data [37].

Given the popularity of AI to complete medical imaging projects that perform organ and lesion detection [38], we predict that segmentation projects will likely see diminishing returns in network performance after a threshold number of data points. As such, large datasets may not be a requirement to performing quality AI imaging research. Study teams can start with smaller datasets and evaluate performance analysis on subsets of the training data to predict the plateau effect in their datasets.

5. Conclusions

The required number of annotated cases for accurate organ segmentation with a deep learning network may be lower than expected. The marginal benefit of more data may diminish after reaching a threshold number of cases in a deep learning network. In this study of prostate organ segmentation, the U-Net CNN plateaued at 160 cases.

Author Contributions: Conceptualization, M.B., D.C., and P.C.; methodology, M.B., and R.H.; software, M.B., C.C., and P.C.; validation, M.B., C.C., and P.C.; formal analysis, M.B., C.C., and P.C.; investigation, M.B., R.H., D.C., and P.C.; resources, D.C., E.U., and P.C.; data curation, M.B., R.H., A.U., J.G.-B., E.U., and P.C.; writing—original

draft preparation, M.B.; writing—review and editing, M.B., R.H., A.U., J.G.-B., M.S., and D.C.; visualization, M.B., R.H., C.C., D.C., and P.C.; supervision, R.H., D.C., and P.C.; project administration, R.H., D.C., and P.C.; funding acquisition, M.B., D.C., and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by a Radiological Society of North America Medical Student Research Grant (RMS1902) and additionally by an Alpha Omega Alpha Carolyn L. Kuckein Student Research Fellowship.

Disclosures: Author Peter Chang, MD, is a cofounder and shareholder of Avicenna.ai, a medical imaging startup. Author Daniel Chow, MD, is a shareholder of Avicenna.ai, a medical imaging startup.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

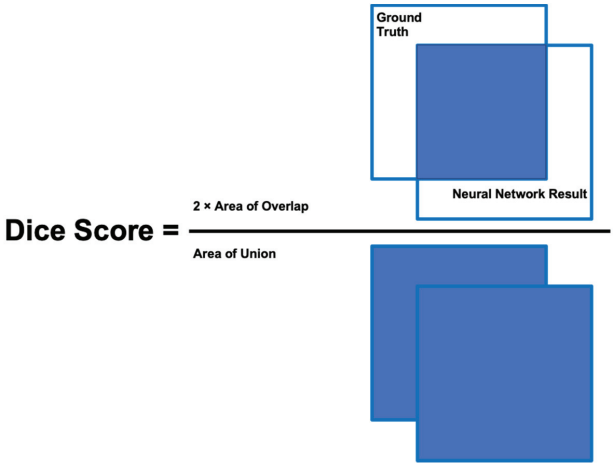


Figure A1. The Dice score is used to measure the performance of the neural network. Its range is from 0 (worst) to 1 (best). A score of 1 demonstrates perfect overlap between the ground truth and the neural network's output. A score of 0 shows that the ground truth and neural network's output have no intersection.

References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
4. Chang, P.; Kuoy, E.; Grinband, J.; Weinberg, B.; Thompson, M.; Homo, R.; Chen, J.; Abcede, H.; Shafie, M.; Sugrue, L. Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. *Am. J. Neuroradiol.* **2018**, *39*, 1609–1616. [CrossRef] [PubMed]
5. Chang, P.; Grinband, J.; Weinberg, B.; Bardis, M.; Khy, M.; Cadena, G.; Su, M.-Y.; Cha, S.; Filippi, C.; Bota, D. Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. *Am. J. Neuroradiol.* **2018**, *39*, 1201–1207. [CrossRef] [PubMed]
6. Figueroa, R.L.; Zeng-Treitler, Q.; Kandula, S.; Ngo, L.H. Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 8. [CrossRef]

7. Fukunaga, K.; Hayes, R.R. Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 873–885. [CrossRef]
8. Gheisari, M.; Wang, G.; Bhuiyan, M.Z.A. A survey on deep learning in big data. In Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017; pp. 173–180.
9. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146–157. [CrossRef]
10. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med. (Lausanne)* **2019**, *6*, 34. [CrossRef]
11. Kohli, M.D.; Summers, R.M.; Geis, J.R. Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session. *J. Digit. Imaging* **2017**, *30*, 392–399. [CrossRef]
12. Baltres, A.; Al Masry, Z.; Zemouri, R.; Valmary-Degano, S.; Arnould, L.; Zerhouni, N.; Devalland, C. Prediction of Oncotype DX recurrence score using deep multi-layer perceptrons in estrogen receptor-positive, HER2-negative breast cancer. *Breast Cancer (Tokyo Jpn.)* **2020**. [CrossRef]
13. Zemouri, R.; Omri, N.; Morello, B.; Devalland, C.; Arnould, L.; Zerhouni, N.; Fnaiech, F. Constructive deep neural network for breast cancer diagnosis. *IFAC-PapersOnLine* **2018**, *51*, 98–103. [CrossRef]
14. Ker, J.; Wang, L.; Rao, J.; Lim, T. Deep learning applications in medical image analysis. *IEEE Access* **2018**, *6*, 9375–9389. [CrossRef]
15. Cho, J.; Lee, K.; Shin, E.; Choy, G.; Do, S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv* **2015**, arXiv:1511.06348.
16. Lakhani, P. Deep convolutional neural networks for endotracheal tube position and X-ray image classification: Challenges and opportunities. *J. Digit. Imaging* **2017**, *30*, 460–468. [CrossRef] [PubMed]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
18. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
19. Zou, K.H.; Warfield, S.K.; Bharatha, A.; Tempany, C.M.; Kaus, M.R.; Haker, S.J.; Wells III, W.M.; Jolesz, F.A.; Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index1: Scientific reports. *Acad. Radiol.* **2004**, *11*, 178–189. [CrossRef]
20. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
21. Clark, T.; Zhang, J.; Baig, S.; Wong, A.; Haider, M.A.; Khalvati, F. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks. *J. Med. Imaging (Bellingham)* **2017**, *4*, 041307. [CrossRef]
22. Rundo, L.; Militello, C.; Russo, G.; Garufi, A.; Vitabile, S.; Gilardi, M.; Mauri, G. Automated prostate gland segmentation based on an unsupervised fuzzy C-means clustering technique using multispectral T1w and T2w MR imaging. *Information* **2017**, *8*, 49. [CrossRef]
23. Ghose, S.; Mitra, J.; Oliver, A.; Marti, R.; Lladó, X.; Freixenet, J.; Vilanova, J.C.; Sidibé, D.; Meriaudeau, F. A random forest based classification approach to prostate segmentation in MRI. *Miccai Grand Chall. Prostate MR Image Segm.* **2012**, *2012*, 125–128.
24. Tian, Z.; Liu, L.; Zhang, Z.; Fei, B. PSNet: Prostate segmentation on MRI based on a convolutional neural network. *J. Med. Imaging* **2018**, *5*, 021208. [CrossRef] [PubMed]
25. Karimi, D.; Samei, G.; Kesch, C.; Nir, G.; Salcudean, S.E. Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1211–1219. [CrossRef] [PubMed]
26. Clark, T.; Wong, A.; Haider, M.A.; Khalvati, F. Fully Deep Convolutional Neural Networks for Segmentation of the Prostate Gland in Diffusion-Weighted MR Images. In Proceedings of the International Conference Image Analysis and Recognition, Montreal, QC, Canada, 5–7 July 2017; pp. 97–104.
27. Zhu, Y.; Wei, R.; Gao, G.; Ding, L.; Zhang, X.; Wang, X.; Zhang, J. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J. Magn. Reson. Imaging* **2019**, *49*, 1149–1156. [CrossRef]

28. Zhu, Q.; Du, B.; Turkbey, B.; Choyke, P.L.; Yan, P. Deeply-supervised CNN for prostate segmentation. In Proceedings of the 2017 International Joint Conference on Neural Networks (Ijcnn), Anchorage, Alaska, 14–19 May 2017; pp. 178–184.
29. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
30. Wang, B.; Lei, Y.; Tian, S.; Wang, T.; Liu, Y.; Patel, P.; Jani, A.B.; Mao, H.; Curran, W.J.; Liu, T. Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. *Med. Phys.* **2019**, *46*, 1707–1718. [CrossRef] [PubMed]
31. Cheng, R.; Roth, H.R.; Lu, L.; Wang, S.; Turkbey, B.; Gandler, W.; McCreedy, E.S.; Agarwal, H.K.; Choyke, P.; Summers, R.M. Active appearance model and deep learning for more accurate prostate segmentation on MRI. In Proceedings of the Medical Imaging 2016: Image Processing, San Diego, CA, USA, 27 February 2016; p. 97842I.
32. Jackson, P.; Hardcastle, N.; Dawe, N.; Kron, T.; Hofman, M.; Hicks, R.J. Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy. *Front. Oncol.* **2018**, *8*, 215. [CrossRef] [PubMed]
33. Dong, X.; Lei, Y.; Wang, T.; Thomas, M.; Tang, L.; Curran, W.J.; Liu, T.; Yang, X. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Med. Phys.* **2019**, *46*, 2157–2168. [CrossRef] [PubMed]
34. Toth, R.; Ribault, J.; Gentile, J.; Sperling, D.; Madabhushi, A. Simultaneous Segmentation of Prostatic Zones Using Active Appearance Models With Multiple Coupled Levelsets. *Comput. Vis. Image Underst.* **2013**, *117*, 1051–1060. [CrossRef]
35. Lay, N.S.; Tsehay, Y.; Greer, M.D.; Turkbey, B.; Kwak, J.T.; Choyke, P.L.; Pinto, P.; Wood, B.J.; Summers, R.M. Detection of prostate cancer in multiparametric MRI using random forest with instance weighting. *J. Med. Imaging* **2017**, *4*, 024506. [CrossRef]
36. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Smolley, S.P. Least squares generative adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2794–2802.
37. Bardis, M.D.; Houshyar, R.; Chang, P.D.; Ushinsky, A.; Glavis-Bloom, J.; Chahine, C.; Bui, T.-L.; Rupasinghe, M.; Filippi, C.G.; Chow, D.S. Applications of Artificial Intelligence to Prostate Multiparametric MRI (mpMRI): Current and Emerging Trends. *Cancers* **2020**, *12*, 1204. [CrossRef]
38. Zemouri, R.; Zerhouni, N.; Racoceanu, D. Deep learning in the biomedical applications: Recent and future status. *Appl. Sci.* **2019**, *9*, 1526. [CrossRef]

electronics

MDPI

Article

Novel Biased Normalized Cuts Approach for the Automatic Segmentation of the Conjunctiva

Giovanni Dimauro [1,*] and Lorenzo Simone [2]

[1] Department of Computer Science, University of Bari, 70125 Bari, Italy;
[2] Department of Computer Science, University of Pisa, 56127 Pisa, Italy; l.simone3@studenti.unipi.it
* Correspondence: giovanni.dimauro@uniba.it

Received: 18 May 2020; Accepted: 9 June 2020; Published: 14 June 2020

Abstract: Anemia is a common public health disease diffused worldwide. In many cases it affects the daily lives of patients needing medical assistance and continuous monitoring. Medical literature states empirical evidence of a correlation between conjunctival pallor on physical examinations and its association with anemia diagnosis. Although humans exhibit a natural expertise in pattern recognition and associative skills based on hue properties, the variance of estimates is high, requiring blood sampling even for monitoring. To design automatic systems for the objective evaluation of pallor utilizing digital images of the conjunctiva, it is necessary to obtain reliable automatic segmentation of the eyelid conjunctiva. In this study, we propose a graph partitioning segmentation approach. The semantic segmentation procedure of a diagnostically meaningful region of interest has been proposed for exploiting normalized cuts for perceptual grouping, thereby introducing a bias towards spectrophotometry features of hemoglobin. The reliability of the identification of the region of interest is demonstrated both with standard metrics and by measuring the correlation between the color of the ROI and the hemoglobin level based on 94 samples distributed in relation to age, sex and hemoglobin concentration. The region of interest automatically segmented is suitable for diagnostic procedures based on quantitative hemoglobin estimation of exposed tissues of the conjunctiva.

Keywords: semantic segmentation; pattern recognition; hemoglobin; anemia; human tissues; conjunctiva; non-invasive medical device

1. Introduction

1.1. Background

Anemia is a blood disorder in which the number of red blood cells is inadequate to carry oxygen to human tissues and organs. It affects about a third of the global population, being the most common blood disorder according to the epidemiological results [1–3]. Each different form of this condition has its specific underlying causes. The process of erythrocyte production in the blood involves bone marrow and erythropoietin, a hormone produced by the kidneys, which regulates the process of erythropoiesis, favoring a constant rate of change in the number of erythrocytes in the blood [4]. Adequate production of red blood cells prevents conditions such as anemia and tissue hypoxia. To promote normal erythropoiesis, correct hemoglobin synthesis is required. Hemoglobin, the iron-containing protein, represents the predominant protein found in erythrocytes, responsible for transporting oxygen from the lungs to the other tissues. Anemia caused by deficiencies of the aforementioned factors results in production patterns of abnormal and different erythrocytes [5]. Diagnosing anemia requires in most cases a complete blood count (CBC) to check different properties, including hemoglobin and hematocrit levels. Each physiological need depends on several factors, such as gender, age, different stages of

pregnancy and altitude. The thresholds presented in Table 1 are used to diagnose anemia in individuals in a screening or clinical setting according to World Health Organization diagnostic guidelines [6].

Table 1. Hemoglobin (Hb) thresholds used to define anemia living at sea level according to the World Health Organization guidelines [6].

Age Group	No Anemia	Mild Anemia	Moderate Anemia	Severe Anemia
Children 5–11 years	$\geq$ 11.5 g/dL	11–11.4 g/dL	8–10.9 g/dL	<8 g/dL
Children 12–14 years	$\geq$ 12 g/dL	11–11.9 g/dL	8–10.9 g/dL	<8 g/dL
Non-pregnant women	$\geq$ 12 g/dL	11–11.9 g/dL	8–10.9 g/dL	<8 g/dL
Pregnant women	$\geq$ 11 g/dL	10–10.9 g/dL	7–9.9 g/dL	<7 g/dL
Men	$\geq$ 13 g/dL	11–12.9 g/dL	8–10.9 g/dL	<8 g/dL

There has always been a worldwide interest in providing simple, cheap and robust procedures to measure hemoglobin without requiring specialized primary health-care workers or medical laboratories [7]. In response to this need, WHO developed the hemoglobin color scale (HCS) in 2001. It consists of a small card of six shades of red from lighter to darker representing a hemoglobin g/dL concentration from 4 to 14 with a step size of 2 g/dL. The specificity of this method has been disputed in literature; for instance, in 2005 14 studies mostly reported a high sensitivity for detecting anemia (75–97%) [8]. Nevertheless, what is crucial about HCS is its potential for opening the way to different approaches requiring a mixture of expertise from different disciplines, such as computer science, in the future. Like other diagnostic-clinical and analytical-laboratory medical disciplines that are beginning to make extensive use of image, sound or signal analysis; and machine and deep learning techniques [9–18], it is worthwhile to invest in research and development of technologies such as those we deal with in this paper, with the dual purpose of significantly reducing the costs borne by the national health systems and powering the healthcare and medical services that would be exempted from a considerable amount of practically useless activities. Since the importance of the objective evaluation of the pallor of the conjunctiva has been understood, a lot has been done. Numerous researchers have worked to develop methods, techniques and devices to make the estimate of the level of hemoglobin or the determination of the condition of severe anemia, in a non-invasive way, as reliable as possible. We will report a summary of this path in the section "Related Works."

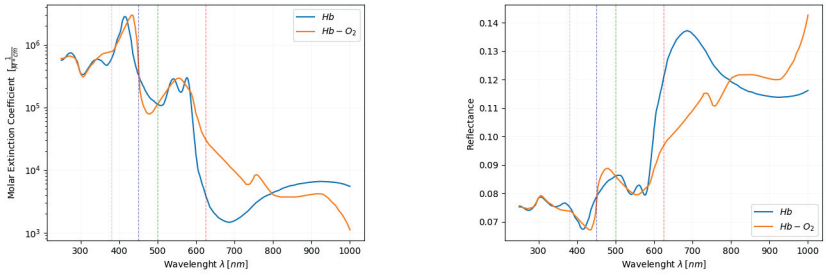
1.2. Haemoglobin Spectrophotometry

HCS and physical examination of exposed tissues such as palpebral conjunctiva or nail beds both rely on how humans perceive colors related to the optical spectrum [19]. To better analyze and handle this phenomenon from a computer vision point of view, a chemical insight is required. Spectrophotometry in chemistry is defined as quantitative measurements of the reflective or absorption properties of a material from a wavelength perspective. The spectra of the hemoglobin molecule vary based on whether it is bound to oxygen, carbon monoxide or nothing; the the latter is also called deoxygenated Hb [20].

We relied on experimental literature data [21] for the absorption spectra of hemoglobin used for both plots in Figure 1. The absorption coefficient μ_a^{Hb} for HbO_2 and Hb is calculated as follows:

$$\mu_a^{Hb}(\lambda) = \frac{2.303 \times e_{Hb}(\lambda)[\frac{L}{cm \times mol}] \times 150[g/L]}{M_{Hb}[g/mol]}, \tag{1}$$

where $e_{Hb}(\lambda)[\frac{L}{cm \times mol}]$ is the Hb molar extinction coefficient and $M_{Hb}[g/mol]$ is the Hb gram molecular weight, assuming a concentration of 150 grams per liter.



(a) **(b)**

Figure 1. Plots visualizing optical absorption and reflectance of Hb and HbO_2, vertical dashed lines are related to human perception of colors associated with (λ). (**a**) Molar extinction coefficient (ϵ) related to absorbance over wavelength (λ) considering 15 g/dL of hemoglobin concentration and 1 cm cuvette. (**b**) Derived reflectance plot of absorbance under same constants.

Over the years, the palpebral conjunctiva has been a good spot to diagnose anemia, representing a highly vascular area characterized by several capillaries. In [22] a multi-layered tissue model is proposed and investigated to approximate the lower eyelid with seven layers: conjunctival epithelium, tarsal plate, orbicularis oculi, subcutaneous tissue, dermis, epidermis and stratum corneum on the outside of the eyelid tissue. The conjunctiva is perfused from the ascending branch of the posterior conjunctival artery. The presence of interweaving capillary networks penetrating several layers of the model, with the mucous membrane being highly transparent, allows for model approximations for the digital image domain. As already visually described by Figure 1, Hb and HbO_2 both absorb wavelengths from 275 to about 550 nm corresponding to a visible spectrum from purple to light green. Each frequency above 600 nm is highly reflected, matching with colors from orange to dark red. A typical human eye is known to be aware of wavelengths in a range from 380 to 740 nm. The cytoplasm of the red blood cell is rich in hemoglobin, that being responsible for the reddish appearance of exposed tissues and blood in general. Laboratory-based experiments conducted in [23,24], inspired us to start from those results to accomplish segmentation and digital image analysis related to hemoglobin.

1.3. Related Works

Over the years many researchers have put in effort toward developing non-invasive methods for anemia detection through hemoglobin estimation. The relevance of conjunctiva hue in the clinical evaluation of anemia was tested in [25] for 219 healthy ambulatory subjects. Three educated non-clinicians, appropriately trained, overall agreed on conjunctiva hue performing with kappa coefficients between 0.27 and 0.34. As a result, hue variation strictly depends on the objective of the assessment and training of field personnel. Comparing earlier results obtained by physical examination and the latest digital photography, the latter is minimizing variance, optimizing specificity and sensitivity by using machine learning and automatic segmentation procedures. Establishing the most successful technology still leaves questions about the best region to analyze exploiting color properties associated with better results. Studies in [26] from an ophthalmology point of view open a debate for correlation of anemia between bulbar conjunctival blood column and palpebral conjunctival hue (PCH). From the results of this study, it seems that the bulbar conjunctiva can be successfully included in the set of interesting features, achieving slightly less specificity than PCH, but higher sensitivity. Paradigms of non-invasive and on-demand diagnostics based on smartphone and digital images are spreading due to the advancing of remote diagnosis and affordability [27–29]. A smartphone camera-based application monitoring blood hemoglobin concentration has been developed in [30]. Utilizing a light source pointed to the patient's finger, they performed a chromatic analysis on 31 samples, achieving

sensitivity and precision of 85.7% and 76.5% respectively; they received Food and Drug Administration agreement. Another smartphone-based self-screening tool is depicted in [31] utilizing fingernail beds digital images. Patients select the regions of interest by themselves, corresponding to the nailbeds, and a result is then displayed on the smartphone screen; camera flash reflections and white spots which may affect Hgb level measurements are removed with a quality control algorithm. They reported an accuracy of $\pm 0.92\ \mathrm{g/dL^{-1}}$ of CBC hemoglobin level with personalized calibration, suggesting the relevance of those systems as a monitoring utility. In our study, we analyzed assumptions from related past works and the clinical correlation between conjunctival pallor and anemia condition [32], proposing a fully automated segmentation algorithm. Throughout this process, color features from hemoglobin reflectance spectrum provide a key role in biasing towards a region of interest proposal.

In the literature, few works deal with the automatic segmentation of the conjunctiva. In particular, reference [33] proposes a method for the automatic segmentation of the palpebral conjunctiva that carries out an image processing process based on the equalization of the image in RGB, filter unsharp masking and red channel masking. In [34] the authors developed an algorithm for automatically segmenting the image by finding a "distinctly red" region, bounded by two parallel long-running edges at the top and the bottom; this is achieved by combining the Canny edge detection technique with morphological operations in the CIELAB color space. However, with the aim of estimating anemia, they stated that their method of segmenting was less reliable than manual conjunctiva segmentation made by an expert physician. In [35] the authors use a threshold triangle (which uses triangle algorithm for thresholding) for binary differentiation between the palpebral conjunctiva and background.

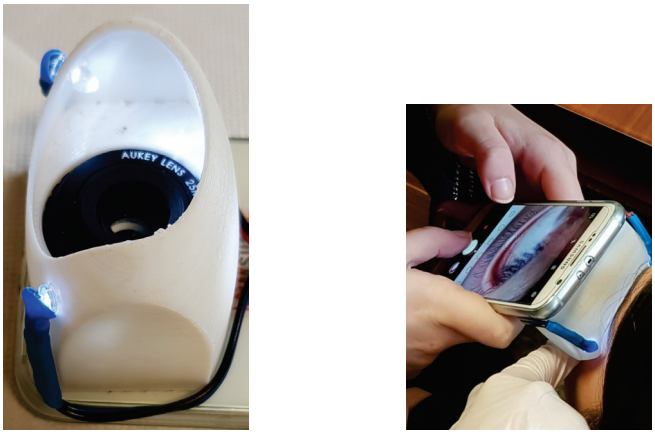
1.4. Image Capturing Methodology

The technique adopted to capture digital images of a patient's conjunctiva was based on the latest approach of a research study conducted in [36–38]. As a recap, the main requirements to designing an effective tool for estimating the condition of anemia through the use of digital images of the palpebral conjunctiva would be:

- Provide an easy to us;e device with affordable hardware components
- Its usage should not require trained medical personnel;
- It should provide remote diagnosis and telemedicine conveniences.

The acquisition system is shown in Figure 2. It consists of a macro-lens assembled into a specially designed, 3D-printed lightened spacer Figure 2a and a typical smartphone as in the real-life application Figure 2b. The lens can take high-resolution images being attached to a smartphone (we used the Aukey PL-M1 25 mm 10x macro lens). The LED lights can be powered directly from the smartphone or a battery applied to the cover of a smartphone. The lens is fixed on the plastic cover of the smartphone: this device allows for obtaining high resolution images close to the eye, insensitive to the ambient lighting conditions.

The dataset used in the present study, which will be described later, has been created with a Samsung S6 smartphone.



(a) **(b)**

Figure 2. (**a**) The acquisition device consists of a special spacer and a macro lens to acquire images with a high-resolution smartphone at close range; (**b**) the moment of the acquisition of an image of the conjunctiva.

2. Proposed Method

Each digital image from the dataset is converted into an RGB color space matrix representation. The segmentation process can be summarized in three different phases: dimensionality reduction by clustering approach, grouping as graph partitioning and a final ROI extraction. The introduction of a preliminary clustering step determines a speed up in N-Cuts performance arising from the theoretical proofs by the N-Cuts original paper regarding computational complexity in terms of both space and time. The algorithm constructing a region adjacency graph (RAG) does not consider each pixel from the original resolution anymore, but groups of them preserving spatial and color differences amongst them. Finally, we aim at grasping a non-linear relation between brightness intensities from the red and green channels, based on previous assumptions of reflectance rate by a spectrophotometry point of view.

2.1. K-Means Dimensionality Reduction

The objective of a clustering task is grouping data instances into subsets maximizing a similarity measure, while different instances should belong to different groups [39–41]. We applied the principles from k-means clustering to image segmentation tasks. The main goal in this phase is to produce a feature space similarly to Voronoi diagrams for planes, reducing the complexity of the graph representing the original image. Each pixel from now on will be referred to as a vector in a five-dimensional space: x and y coordinates from the matrix; R, G and B channel intensities from color representation.

$$f(x,y) = \overrightarrow{p} = \alpha_x \overrightarrow{p_x} + \alpha_y \overrightarrow{p_y} + \alpha_r \overrightarrow{p_r} + \alpha_g \overrightarrow{p_g} + \alpha_b \overrightarrow{p_b} \tag{2}$$

This approach allows us to iteratively minimize the sum of distances from each pixel to its cluster centroid. We briefly summarize the steps of the algorithm as follows:

1. Initialize centroid vectors.
2. Pixels retain spatial as well as color features, allowing us to define an appropriate weighted Euclidean distance as a measure of similarity between them. For each of them, calculate the distance d between the centroid and each pixel of the image defined as:

$$d(\overrightarrow{u}, \overrightarrow{v}) = \| \overrightarrow{u} - \overrightarrow{v} \| = \sqrt{(u_x - v_x)^2 + (u_y - v_y)^2 + (u_r - v_r)^2 + (u_g - v_g)^2 + (u_b - v_b)^2} \quad (3)$$

3. Each pixel is assigned to the centroid minimizing d.
4. Recalculate the position of each centroid c_k where $\overrightarrow{p_{ki}}$ is the i_{th} pixel contained in k_{th} centroid using the relation:

$$c_k = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{p_{ki}} \quad (4)$$

This approach included in the broader field of unsupervised learning approaches, consists of initial batch updates, in which at each step we reassign points to their nearest cluster centroid, followed by cluster centroid recalculations. In online updates, the points are reassigned only if reducing the sum of intra-cluster distances. Those updates already converge towards a local minimum in short order.

In Figure 3, the original image is processed with a three-dimensional (R, G and B) space and in the last picture with a five-dimensional model including both color and spatial features. In the latter, there is not an increase in computational complexity since the only calculation affected is the distance function. However, in each digital image analyzed, the intra-cluster variance is minimized efficiently with properly outlined boundaries in between each group of pixels. The classified instances closer to mucocutaneous junction are noisy in the first approach, while on the second one each semantic class (iris, pupil, sclera, eyelid, and conjunctiva) appears as a compact union of clusters.

(a) **(b)** **(c)**

Figure 3. (**a**) Original digital image acquired; (**b**) k-means clustering procedure using only three dimensional (R,G and B) channels from color space; (**c**) proposed k-means procedure with a model in five dimensions retaining both spatial and color properties.

2.2. Normalized Cuts Segmentation

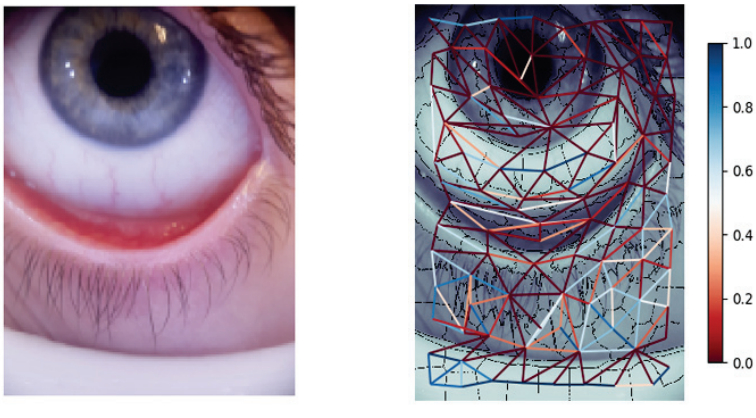
K-means as a clustering algorithm is a valuable approach for exploiting local impressions of a scene, but it lacks in providing a global or hierarchical perspective. For this reason, we take advantage of a grouping algorithm treating the segmentation task as a graph partitioning problem, such as NCuts. It has a better ability to generalize when applied to different scenarios. Conventionally, the normalized cut is an unbiased measure of dissimilarity between graph subgroups [42]. We have converted the set of superpixels from a five-dimensional feature space in a weighted undirected graph $G = (V, E)$. Each point is included in the set of nodes having one edge for each pair of vertices.

The region adjacency graph is constructed based on precomputed areas from the k-means segmentation algorithm. Each connection amongst them is depicted in Figure 4b and representable in a weight matrix W. The edge weight w_{ij} from node i to node j is defined as in the standard approach of normalized cuts as a product of a feature similarity and a spatial term. $X(i)$ is the coordinate vector of the centroid pixel and $F(i)$ is a feature vector based on averaged R, G and B intensities of each pixel in the area. The value r acts as a proximity threshold based on the Euclidean distances amongst precomputed centroids. In our specific application we have tried different configurations ranging from 3 to 100, regulating the sparsity of the weight matrix but not impacting the segmentation outcome. Weights and features are described by the following equations:

$$w_{i,j} = e^{-\frac{\|F(i)-F(j)\|_2^2}{\sigma_I}} * \begin{cases} e^{\frac{-\|X(i)-X(j)\|_2^2}{\sigma_X}}, & \text{if } \|X(i) - X(j)\|_2 < r \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$F(i) = \left[\frac{1}{n}\sum_{j=1}^{n} p_{jr} \quad \frac{1}{n}\sum_{j=1} np_{jg} \quad \frac{1}{n}\sum_{j=1} np_{jb}\right] \tag{6}$$

The algorithm is capable of extracting significant components from each sample from the dataset, avoiding intra-cluster variations.



(a) **(b)**

Figure 4. (**a**) Acquired sample; (**b**) region adjacency graph (RAG) displaying a measure of similarity between each region. The center of each node is considered a vertex. For each connection between two regions, there is an associated colored line according to the measure of similarity.

In Figure 5, we added a visual semantic description of the resulting cuts. With this phase, we raise the level of abstraction of the segmentation, starting from the clusters of Figure 3; we end up with features closer to an anatomical perspective. The small gap in colors between the conjunctival area and mucocutaneous junction is perfectly delineated in each sample from the dataset, paving the way for a machine-learning-based anemia estimator.

In the proposed segmentation output from Figure 5, a recursive approach could be run to further decompose regions of interest from the conjunctival area. As an example, this could lead to a better parting of the two conjunctivae, palpebral and forniceal, so as to contribute to the open debate about the prevalence of one or the other as the best estimator of anemia [43]. In fact, the palpebral conjunctiva highlights the vascularization of the underlying area better than the forniceal and probably allows highlighting minimal variations of blood color. The assumption seems confirmed by scientific literature. However, some authors take into consideration the whole conjunctiva, including both palpebral

and forniceal, to construct and validate their models. It is still an open problem. Furthermore, in [43] the authors state that it should be interesting to establish whether the investigations carried out on a small portion of the conjunctiva can be sufficient and position independent. In fact, the sparsity and density of the blood micro-vessels can change in different parts of the eyelid. Therefore, the recursive identification of further clusters can help to answer the above questions.

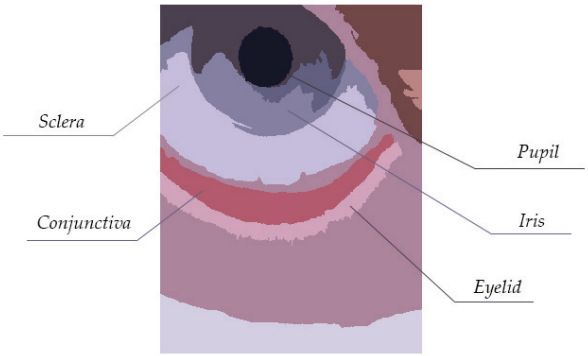


Figure 5. Segmentation output result with semantic class description of eye anatomy.

2.3. Hemoglobin Heatmap Coefficients

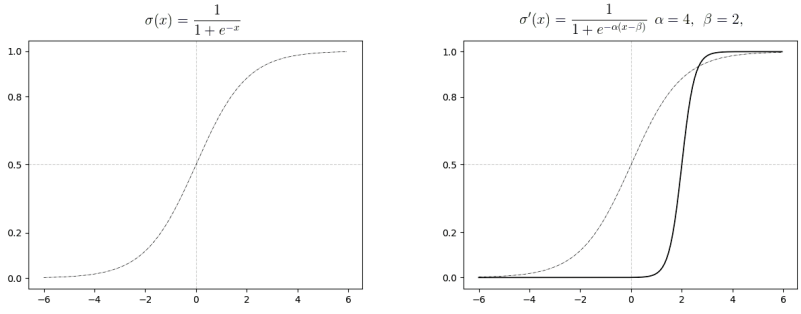
In medical image or radar signal processing tasks, contrast enhancement is a widely used technique in various applications, ranging from improving the quality of photographs acquired in poor conditions [44] to emphasizing regions of interest [45,46]. Histogram equalization is one of the most common approach due to its simple mechanism and effectiveness, but as a drawback, image brightness usually changes after the procedure, caused by its flattening behavior. In our study the objective is focused on approximating the spectrophotometry multi-layered reflectance model investigated in Section 1.2, grasping a mathematical description for digital images. In the literature several studies apply spectral domain scanning, resulting in a time-consuming acquisition process and expensive equipment. This approach does not fit our needs of developing a cheap, non-invasive diagnostic tool. An example of an ill-posed problem known as spectral reconstruction from an RGB scene has been conducted with deep learning techniques in [47,48]. Lastly, researches are highly promoting the validity of these approaches, but despite this, our application domain allows us to further reduce the solution required. Our method, interpreting the image as a signal, performs a pixel pointwise non-linear transformation from red and green color space values, returning a coefficient highlighting vascularized regions. In the literature, the ratio between R and G channels has often been used as a guide to spot those areas, thereby finding the highest values in forniceal and palpebral conjunctival tissues. We propose a generalized logistic function filtering technique including more flexibility than a standard sigmoid. Considering an image I as a vector in three channel functions based on grid coordinates, we obtain the following σ' transformation:

$$I(x,y) = \begin{bmatrix} r(x,y) \\ g(x,y) \\ b(x,y) \end{bmatrix}, \qquad \sigma'(I,x,y) = \frac{1}{1+e^{-\alpha\left(\frac{I_r(x,y)}{I_g(x,y)}-\beta\right)}} \tag{7}$$

The parameter α determines the slope of the function, emphasizing the discrepancy in terms of ratio between color channels; β acts as a minimum ratio threshold for the activation of each pixel.

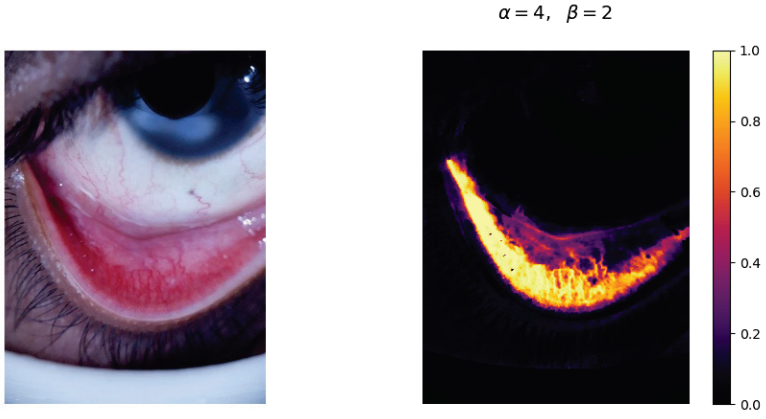
A comparison of the behavior of standard and generalized logistic function with parameterization $\alpha = 4$ and $\beta = 2$ is depicted in Figure 6. This parameterization yielded results with a remarkable

capability of generalizing well in diagnostic imaging ranging from conjunctival tissue to endoscopic domains. Increasing values of α related to the steepness, tend towards the trivial case of a binarization step function losing information about the relationship underlying a variety of brightness ratios. An application of this model is illustrated in Figure 7 useful for digital images of the conjunctival region.



(a) **(b)**

Figure 6. (**a**) Standard logistic function plot. (**b**) Generalized logistic function plot using parameters $\alpha = 4$ and $\beta = 2$.



(a) **(b)**

Figure 7. (**a**) Acquired sample. (**b**) Heatmap plot of the scoring matrix displaying the magnitudes of the coefficients computed by applying the generalized sigmoid function on the acquired sample.

The real values range from 0 to 1 according to σ' function definition. The filtering process produces a scoring matrix assigning lower values to the background, including the sclera, pupil, iris, eyelid and white support platform from the device. Palpebral and forniceal conjunctiva are primarily perfused by both internal and external carotid arteries; this is reflected in high values from the scoring matrix ranging from 0.7 to 1, and the respective blood vessels are significantly highlighted, as shown in Figure 7b.

Since we are interested in obtaining a semantic interpretation out of the regions proposed by NCut, the matrix of coefficients acts as an effective bias for calculating the probability distribution of each class. Edge weights crossed by aggregated pixels resulting from σ' are strengthened or decreased, resulting

in a region proposal based on the magnitude of the connection. In Figure 8, we provide a subset of 10 digital images from the dataset, showing the qualitative difference between the proposed semantic segmentation (top row) and the manually segmented ground truth (second row). In Figure 9 we provide two samples of erroneous acquisitions in order to show the robustness of the proposed segmentation in unusual conditions; in fact, only images with excellent characteristics can provide useful information for the correct estimation of anemia.

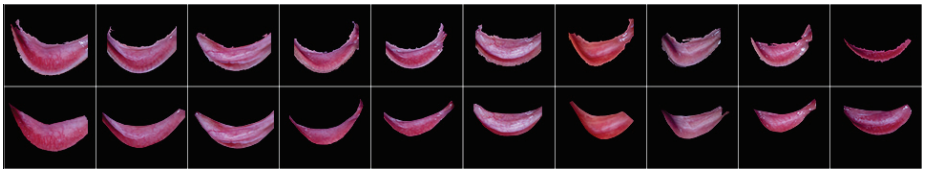


Figure 8. The top row represents a subset of samples automatically segmented with the proposed approach. The ordered second row depicts the mapping with the manual segmentation ground truth of the conjunctival region.

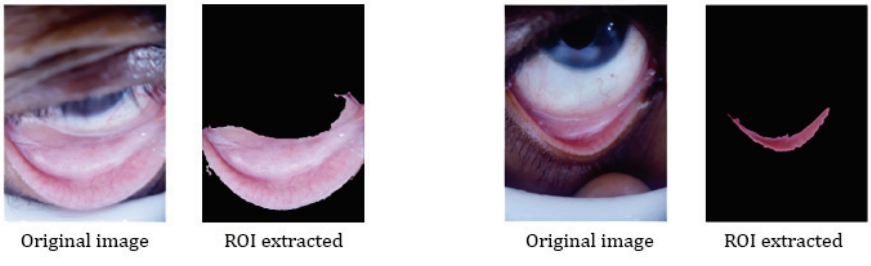


Figure 9. Examples of two images that would normally be discarded: the first because the eyelid overlaps the edge of the white spacer and is not perfectly in focus; additionally, the second one is not in focus and the finger appears to lower the eyelid. In both cases, automatic segmentation would still provide an acceptable result.

3. Results

The digital images of the patients' eyes have been captured by the device reported in Figure 2 and assembled on a Samsung S6 smartphone; 94 patients were involved, aged 19–75 (average 34), 46 female and 48 male, with Hb level concentrations in the range of 7.6–17.1 g/dL (average of 11.45 g/dL).

Each picture underwent a manual selection process, isolating and cropping regions of palpebral and forniceal conjunctiva, as shown in Figure 10. This step is needed to compare the manually segmented images considered as the ground truth with the automatic segmentation output from the proposed model. We evaluated both spatial and color properties of regions of interest by assessing the most suitable metrics based on this specific medical image segmentation problem [49]. F1 (FMS1), also known as the Sørensen–Dice coefficient, is the harmonic mean of precision and recall, defined as follows for binary segmentation applications:

$$F_1 = 2 \cdot \left(\frac{Precision \cdot Recall}{Precision + Recall}\right) = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{8}$$

(a) **(b)** **(c)**

Figure 10. (**a**) Manually segmented conjunctiva used as ground truth. (**b**) Automatically segmented conjunctiva obtained by the proposed approach. (**c**) Visualization of the overlapping between green ground truth image and white automatically segmented image ($F1 = 0.904$, $accuracy = 96.41\%$).

The Dice coefficient being an overlapping measure ranging from 0 to 1, gives us a useful perspective about the quality of the segmentation. We are also interested in a calculation involving the number of pixels classified as non-relevant (false positive rate), which is not taken into account either by Dice coefficient or by Jaccard similarity. Accuracy metric is helpful in this case by outlining the rate of correctly classified pixels over the full image.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

With the aim of assessing an average for the overlapping metrics, we computed a binary confusion matrix for each image. The values of this matrix refer to the number of pixels linked to set intersection or set difference between ground truth image and proposed segmentation, which are visually described in Figure 10c.

The averaged summation of each confusion matrix is summarized in Table 2. To give the reader the opportunity to observe the indicators for each sample included in the dataset, in Table A1 we have reported the values of the above metrics in a complete manner. Higher values of specificity for this segmentation task highlight the eligibility to disregard non conjunctival regions with proper confidence. On the other hand, sensitivity as well as F1 being overlapping measures, can reasonably fluctuate with higher variance, meaning in most cases that a finer meaningful subset of the conjunctival region has been selected.

Table 2. Metrics of averaged results of the comparison between manually and automatically segmented images of the conjunctiva.

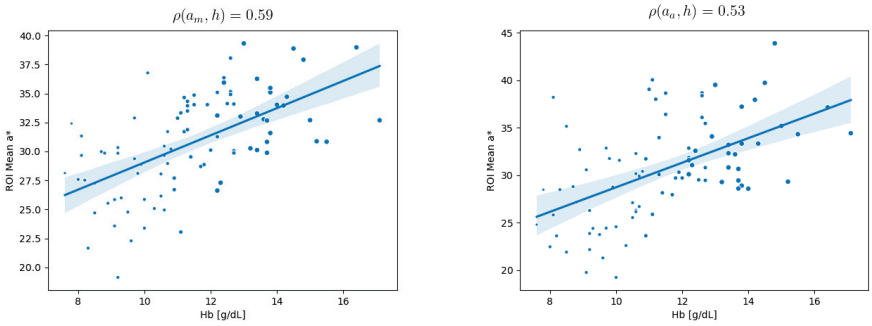
	F1-Measure	**Accuracy**	**Sensitivity (TPR)**	**Specificity (TNR)**
Predicted ROIs	0.7363	93.79%	86.73%	94.63%

The optimal results indicated by the above metrics are sufficient to state the effectiveness of our segmentation algorithm. Since here we are dealing with a rigorous diagnostic procedure, if comparing the precision of the overlapping between proposed and ground truth ROIs is acceptable, we think that a further investigation of the color properties for left-out or added regions would be interesting.

CIELAB is one of the most useful amongst color spaces for erythema analysis and computer vision for diagnostics, composed by an approximately uniform three-dimensional space: L*, a*, b*. A widely used dimension from this space, a*, has a well-known correlation with hemoglobin values in this domain [36–38]. Our purpose is to examine the strength of linear correlation between mean

values of a* extracted from digital images of conjunctivas and the relative Hb g/dL concentration from blood samples taken almost at the same time of picture capturing phase (Figure 11). Generalizing the idea of Pearson correlation coefficient (PCC) from two random variables to two standardized vectors, we can estimate the weight of their linear correlation ranging from −1 to 1 and defined by the following equation:

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{a_i - \mu_a}{\sigma_a}\right) \cdot \left(\frac{b_i - \mu_b}{\sigma_b}\right) \tag{10}$$



(a) **(b)**

Figure 11. (**a**) Linear regression and strength of correlation between a* from manual segmentation and Hb g/dL standardized vectors. (**b**) Linear regression and strength of correlation between a* from automatic segmentation and Hb g/dL standardized vectors.

We computed PCC between the mean a* values for both manually and automatically segmented images and Hb g/dL through the entire dataset of 94 samples, thereby obtaining respectively 0.59 and 0.53. The results reconfirm not only the moderate linear correlation between those values, but also a robust contiguity among human based manual segmentation and fully automated segmentation approach proposed.

4. Conclusions

We developed a fully automated segmentation procedure, based on graph partitioning, that exposes conjunctival regions while maximizing the correlation between color properties and hemoglobin concentration in the blood, according to the multi-layered anatomical structures of these tissues. The ROIs extracted by the model underwent an in-depth quantitative comparison with ground truth, using state of the art metrics for similarity and PCC between the a* component from CIELAB space and hemoglobin values. The results attest to the reliability and the capability of generalizing between patients belonging to heterogeneous classes, as the accuracy of the overlap between the manual and automatic ROIs selections, measured with classic metrics, is very good, and the correlation obtained between the level of Hb measured in vivo and that estimated through the color of the manual/automatic ROI are comparable. The proposed method paves the way for further studies involving deep learning techniques for both classifications of an estimated anemia risk category and regression to predict Hb real values. With this study we contribute to the broader diagnostic research field of image processing and analysis of the conjunctival pallor related to anemia diagnosis support. The advancement provided to this non-invasive image capturing procedure will lead to the possibility of embedding the model in a wearable device screening Hb risk category in real-time, without the need for physician support.

Author Contributions: The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Results computed from the confusion matrices of the comparison between manually and automatically segmented images of the conjunctiva for the entire dataset of 94 samples.

Image ID	F1-Measure	Accuracy	Sensitivity (TPR)	Specificity (TNR)
164733	0.7547	0.9097	0.6060	1.0000
918410	0.7647	0.9287	0.6369	0.9567
094523	0.9123	0.9674	0.8696	0.9595
103722	0.6429	0.9687	0.6103	0.6792
190841	0.7011	0.8625	0.586	0.8724
154215	0.6494	0.9558	0.5505	0.7915
160737	0.7844	0.9327	0.6470	0.9957
155221	0.7179	0.9813	0.7044	0.7319
122613	0.7616	0.9176	0.8953	0.6627
132714	0.6641	0.8779	0.4971	1.0000
140525	0.7250	0.9316	0.9255	0.5959
154320	0.5296	0.8965	0.3602	1.0000
143315	0.7563	0.8955	0.6081	1.0000
145200	0.7834	0.9837	0.7677	0.7997
150240	0.6542	0.9170	0.4861	1.0000
155237	0.7672	0.9549	0.9374	0.6493
801000	0.7595	0.9460	0.9613	0.6277
121216	0.7848	0.9521	0.6534	0.9823
120556	0.6804	0.9080	0.5207	0.9815
134128	0.7827	0.9675	0.6715	0.938
150536	0.8229	0.9769	0.8237	0.8221
151234	0.7343	0.9285	0.6025	0.9400
155418	0.8351	0.9757	0.8186	0.8523
152136	0.7407	0.9264	0.8862	0.6362
152924	0.6875	0.9653	0.5282	0.9846
153536	0.6818	0.8958	0.5174	0.9995
154129	0.8665	0.9596	0.9719	0.7817
154759	0.8436	0.9559	0.7770	0.9226
155456	0.8463	0.9539	0.8111	0.8846
160045	0.6242	0.9333	0.4544	0.9965
123002	0.7943	0.9244	0.6703	0.9745
122915	0.7728	0.9664	0.7984	0.7488
232040	0.6222	0.9300	0.5065	0.8064
160522	0.8019	0.9790	0.8133	0.7909
121836	0.5998	0.8646	0.5157	0.7166
134745	0.7401	0.8944	0.7800	0.7040
211040	0.4881	0.9146	0.3258	0.9724
210631	0.9184	0.9838	0.9235	0.9134
223744	0.7676	0.9013	0.6468	0.9440

Table A1. *Cont.*

Image ID	F1-Measure	Accuracy	Sensitivity (TPR)	Specificity (TNR)
224452	0.655	0.8827	0.4872	0.9991
231923	0.7167	0.9513	0.5585	0.9999
232931	0.8046	0.9636	0.7029	0.9406
141804	0.7793	0.9310	0.7248	0.8428
152107	0.6693	0.9144	0.5063	0.9871
161452	0.7892	0.8955	0.7651	0.9673
154641	0.8193	0.9806	0.8627	0.7801
210419	0.8587	0.9675	0.8427	0.8753
221400	0.8056	0.9256	0.6767	0.9952
222325	0.8093	0.9298	0.6913	0.9758
140311	0.6237	0.9594	0.4608	0.9645
180148	0.8293	0.9154	0.7085	0.9998
183506	0.7559	0.9214	0.6226	0.9617
195511	0.7103	0.9149	0.5554	0.9849
201501	0.7197	0.9031	0.5662	0.9874
184029	0.7589	0.9715	0.6305	0.9531
184734	0.8508	0.9636	0.8814	0.8221
185602	0.8863	0.9722	0.8574	0.9172
190638	0.8229	0.9267	0.7120	0.9747
191233	0.8163	0.9388	0.8559	0.7801
191620	0.6685	0.8737	0.7922	0.5782
194457	0.7283	0.9508	0.5858	0.9624
114700	0.6357	0.9133	0.5007	0.8705
115146	0.6255	0.8800	0.5202	0.7842
115853	0.8018	0.9526	0.7490	0.8626
120426	0.6434	0.9588	0.5084	0.8762
202058	0.6903	0.8737	0.5271	1.0000
123714	0.7709	0.9415	0.8038	0.7406
133633	0.6015	0.9539	0.4604	0.8673
143301	0.8145	0.9803	0.7065	0.9614
144551	0.7174	0.9540	0.8865	0.6025
145301	0.6573	0.9124	0.4972	0.9693
150804	0.6424	0.9447	0.4849	0.9515
150539	0.8357	0.9547	0.7311	0.9750
151450	0.7388	0.9020	0.5886	0.9917
153146	0.7744	0.9295	0.6382	0.9844
162916	0.7940	0.9369	0.6713	0.9716
202947	0.9040	0.9641	0.8552	0.9587
180925	0.7136	0.9124	0.6152	0.8494
190130	0.8209	0.9776	0.7666	0.8834
190334	0.6594	0.9354	0.7855	0.5682
121621	0.8401	0.9549	0.7570	0.9436
154729	0.4816	0.9293	0.3244	0.9343
205012	0.8539	0.9651	0.9005	0.8120
205445	0.8337	0.9887	0.8632	0.8063
222551	0.7993	0.9394	0.7278	0.8863
223503	0.8563	0.9834	0.8353	0.8783

Table A1. *Cont.*

Image ID	F1-Measure	Accuracy	Sensitivity (TPR)	Specificity (TNR)
224240	0.7352	0.9379	0.6334	0.8760
205917	0.7118	0.9691	0.6264	0.8242
225922	0.7938	0.9492	0.8498	0.7447
231050	0.7480	0.9386	0.7003	0.8027
183626	0.5987	0.9463	0.4453	0.9133
161347	0.7855	0.9466	0.7371	0.8406
130148	0.6814	0.9690	0.5243	0.9728
130225	0.7896	0.9383	0.6632	0.9757

References

1. World Health Organization. *Worldwide Prevalence of Anaemia 1993–2005 : WHO Global Database on Anaemia*; de Benoist, B., McLean, E., Egli, I., Cogswell, M., Eds.; WHO: Geneva, Switzerland, 2008.
2. World Health Organization. *The World Health Report 2002*; World Health Organization: Geneva, Switzerland, 2002.
3. McLean, E.; Cogswell, M.; Egli, I.; Wojdyla, D.; Benoist, B. Worldwide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993–2005. *Public Health Nutr.* **2008**, *12*, 444–54. [CrossRef] [PubMed]
4. Koury, M.J. Red blood cell production and kinetics. In *Rossi's Principles of Transfusion Medicine*; Wiley: Hoboken, NJ, USA, 2016, pp. 85–96. [CrossRef]
5. White, J.; Porwit, A.M. *Blood and Bone Marrow Pathology*; Elsevier: Amsterdam, The Netherlands, 2011.
6. World Health Organization; Centers for Disease Control and Prevention. *Assessing the Iron Status of Populations*; World Health Organization, Department of Nutrition for Health and Development: Geneva, Switzerland, 2005.
7. Marn, H.; Critchley, J.A. Accuracy of the WHO Haemoglobin Colour Scale for the diagnosis of anaemia in primary health care settings in low-income countries: A systematic review and meta-analysis. *Lancet Glob. Health* **2016**, *4*, e251–e265. [CrossRef]
8. Critchley, J.; Bates, I. Haemoglobin colour scale for anaemia diagnosis where there is no laboratory: A systematic review. *Int. J. Epidemiol.* **2005**, *34*, 1425–1434. [CrossRef] [PubMed]
9. Dimauro, G.; Girardi, F.; Gelardi, M.; Bevilacqua, V.; Caivano, D. Rhino-Cyt: A System for Supporting the Rhinologist in the Analysis of Nasal Cytology. *Intell. Comput. Theor. Appl. Lect. Notes Comput. Sci.* **2018**, 619–630. [CrossRef]
10. Dimauro, G.; Ciprandi, G.; Deperte, F.; Girardi, F.; Ladisa, E.; Latrofa, S.; Gelardi, M. Nasal cytology with deep learning techniques. *Int. J. Med Informatics* **2019**, *122*, 13–19. [CrossRef] [PubMed]
11. Triggiani, A.; Bevilacqua, V.; Brunetti, A.; Lizio, R.; Tattoli, G.; Cassano, F.; Soricelli, A.; Ferri, R.; Nobili, F.; Gesualdo, L.; et al. Classification of healthy subjects and Alzheimer's disease patients with dementia from cortical sources of resting state EEG rhythms: A study using artificial neural networks. *Front. Neurosci.* **2017**, *10*. [CrossRef]
12. Bevilacqua, V.; Pannarale, P.; Abbrescia, M.; Cava, C.; Paradiso, A.; Tommasi, S. Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression. *BMC Bioinform.* **2012**, *13*. [CrossRef]
13. Bevilacqua, V.; Cariello, L.; Columbo, D.; Daleno, D.; Fabiano, M.D.; Giannini, M.; Mastronardi, G.; Castellano, M. Retinal fundus biometric analysis for personal identifications. In Proceedings of the International Conference on Intelligent Computing, Shanghai, China, 5–18 September 2008; Springer: Berlin, Germany, 2008; pp. 1229–1237.
14. Bevilacqua, V.; D'Ambruoso, D.; Mandolino, G.; Suma, M. A new tool to support diagnosis of neurological disorders by means of facial expressions. In Proceedings of the IEEE International Symposium on Medical Measurements and Applications, Bari, Italy, 30–31 May 2011; pp. 544–549.

15. Dimauro, G.; Caivano, D.; Bevilacqua, V.; Girardi, F.; Napoletano, V. VoxTester, software for digital evaluation of speech changes in Parkinson disease. In Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MeMeA), Benevento, Italy, 15–18 May 2016. [CrossRef]
16. Bevilacqua, V.; Brunetti, A.; Trotta, G.F.; Dimauro, G.; Elez, K.; Alberotanza, V.; Scardapane, A. A novel approach for Hepatocellular Carcinoma detection and classification based on triphasic CT Protocol. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), San Sebastian, Spain, 5–8 June 2017. [CrossRef]
17. Dimauro, G.; Nicola, V.D.; Bevilacqua, V.; Caivano, D.; Girardi, F. Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System. *IEEE Access* **2017**, *5*, 22199–22208. [CrossRef]
18. Dimauro, G.; Caivano, D.; Girardi, F.; Ciccone, M.M. The patient centered Electronic Multimedia Health Fascicle-EMHF. In Proceedings of the IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), Rome, Italy, 17 October 2014. [CrossRef]
19. Collings, S.; Thompson, O.; Hirst, E.; Goossens, L.; George, A.; Weinkove, R. Non-Invasive Detection of Anaemia Using Digital Photographs of the Conjunctiva. *PLoS ONE* **2016**, *11*, e0153286. [CrossRef]
20. Townsend, D.; D'Aiuto, F.; Deanfield, J. Super actinic 420 nm light-emitting diodes for estimating relative microvascular hemoglobin oxygen saturation. *J. Med Biol. Eng.* **2014**, *34*, 172–177. [CrossRef]
21. Zhao, Y.; Qiu, L.; Sun, Y.; Huang, C.; Li, T. Optimal hemoglobin extinction coefficient data set for near-infrared spectroscopy. *Biomed. Opt. Express* **2017**, *8*, 5151. [CrossRef] [PubMed]
22. Kim, O.; McMurdy, J.; Jay, G.; Lines, C.; Crawford, G.; Alber, M. Combined reflectance spectroscopy and stochastic modeling approach for noninvasive hemoglobin determination via palpebral conjunctiva. *Physiol. Rep.* **2014**, *2*, e00192. [CrossRef] [PubMed]
23. Sengupta, B. Biophysical Characterization of Genistein in Its Natural Carrier Human Hemoglobin Using Spectroscopic and Computational Approaches. *Food Nutr.* **2013**, *4*, 83–92.
24. Horecker, B. The absorption spectra of hemoglobin and its derivatives in the visible and near infra-red regions. *J. Biol. Chem.* **1943**, *148*, 173–183.
25. Sanchez-Carrillo, C. Bias due to conjunctiva hue and the clinical assessment of anemia. *J. Clin. Epidemiol.* **1989**, *42*, 751–754. [CrossRef]
26. Kent, A.; Elsing, S.; Hebert, R. Conjunctival vasculature in the assessment of anemia. *Ophthalmology* **2000**, *107*, 274–277. [CrossRef]
27. Kanchi, S.; Sabela, M.I.; Mdluli, P.S.; Inamuddin.; Bisetty, K. Smartphone based bioanalytical and diagnosis applications: A review. *Biosens. Bioelectron.* **2018**, *102*, 136–149. [CrossRef]
28. Escobedo, P.; Palma, A.J.; Erenas, M.M.; Olmos, A.M.; Carvajal, M.A.; Chavez, M.T.; Gonzalez, M.A.L.; Diaz-Mochon, J.J.; Pernagallo, S.; Capitan-Vallvey, L.F.; et al. Smartphone-Based Diagnosis of Parasitic Infections With Colorimetric Assays in Centrifuge Tubes. *IEEE Access* **2019**, *7*, 185677–185686. [CrossRef]
29. Ogirala, T.; Eapen, A.; Salvante, K.G.; Rapaport, T.; Nepomnaschy, P.A.; Parameswaran, A.M. Smartphone-based colorimetric ELISA implementation for determination of women's reproductive steroid hormone profiles. *Med Biol. Eng. Comput.* **2017**, *55*, 1735–1741. [CrossRef] [PubMed]
30. Wang, E.; Li, W.; Hawkins, D.; Gernsheimer, T.; Norby-Slycord, C.; Patel, S. HemaApp: Noninvasive Blood Screening of Hemoglobin Using Smartphone Cameras. *Getmobile: Mob. Comput. Commun.* **2017**, *21*, 26–30. [CrossRef]
31. Mannino, R.; Myers, D.; Tyburski, E.; Caruso, C.; Boudreaux, J.; Leong, T.; Clifford, G.; Lam, W. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nat. Commun.* **2018**, *9*. [CrossRef]
32. Sheth, T.; Choudhry, N.; Bowes, M.; Detsky, A. The Relation of Conjunctival Pallor to the Presence of Anemia. *J. Gen. Intern. Med.* **1997**, *12*, 102–106. [CrossRef] [PubMed]
33. Delgado-Rivera, G.; Roman-Gonzalez, A.; Alva-Mantari, A.; Saldivar-Espinoza, B.; Zimic, M.; Barrientos-Porras, F.; Salguedo-Bohorquez, M. Method for the Automatic Segmentation of the Palpebral Conjunctiva using Image Processing. In Proceedings of the IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA), Concepcion, Chile, 17–19 Ocotber 2018; pp. 1–4.
34. Bevilacqua, V.; Dimauro, G.; Marino, F.; Brunetti, A.; Cassano, F.; Maio, A.D.; Nasca, E.; Trotta, G.F.; Girardi, F.; Ostuni, A.; et al. A novel approach to evaluate blood parameters using computer vision techniques. In Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MeMeA), Benevento, Italy, 12–14 May 2016. [CrossRef]

35. Bauskar, S.; Jain, P.; Gyanchandani, M. A Noninvasive Computerized Technique to Detect Anemia Using Images of Eye Conjunctiva. *Pattern Recognit. Image Anal.* **2019**, *29*, 438–446. [CrossRef]
36. Dimauro, G.; Caivano, D.; Girardi, F. A new method and a non-invasive device to estimate anaemia based on digital images of the conjunctiva. *IEEE Access* **2018**, 1. [CrossRef]
37. Dimauro, G.; Guarini, A.; Caivano, D.; Girardi, F.; Pasciolla, C.; Iacobazzi, A. Detecting clinical signs of anaemia from digital images of the palpebral conjunctiva. *IEEE Access* **2019**, 1. [CrossRef]
38. Dimauro, G.; Baldari, L.; Caivano, D.; Colucci, G.; Girardi, F. Automatic Segmentation of Relevant Sections of the Conjunctiva for Non-Invasive Anemia Detection. In Proceedings of the 3rd International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 26–29 June 2018; pp. 1–5.
39. Dhanachandra, N.; Manglem, K.; Chanu, Y.J. Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Comput. Sci.* **2015**, *54*, 764–771. [CrossRef]
40. Wu, M.N.; Lin, C.C.; Chang, C.C. Brain Tumor Detection Using Color-Based K-Means Clustering Segmentation. In Proceedings of the Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007), Kaohsiung, Taiwan, 26–28 November 2007. [CrossRef]
41. Chitade, A.; Katiyar, S. Color based image segmentation using K-means clustering. *Int. J. Eng. Sci. Technol.* **2010**, *2*, 5319–5325.
42. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *22*. [CrossRef]
43. Dimauro, G.; De Ruvo, S.; Di Terlizzi, F.; Ruggieri, A.; Volpe, V.; Colizzi, L.; Girardi, F. Estimate of Anemia with New Non-Invasive Systems—A Moment of Reflection. *Electronics* **2020**, *9*, 780. [CrossRef]
44. Tan, K.; Oakley, J. Enhancement Of Color Images In Poor Visibility Conditions. In Proceedings of the ICIP International Conference on Image Processing, Vancouver, BC, Canada, 10–13 September 2000. [CrossRef]
45. Arce, G.R.; Bacca, J.; Paredes, J.L. Nonlinear Filtering for Image Analysis and Enhancement. *Essent. Guide Image Process.* **2009**, 263–291. [CrossRef]
46. Graif, M.; Bydder, G.M.; Steiner, R.E.; Niendorf, P.; Thomas, D.; Young, I.R. Contrast-enhanced MR imaging of malignant brain tumors. *Am. J. Neuroradiol.* **1985**, *6*, 855–862. [PubMed]
47. Mammography, O.; Laine, A.; Fan, J.; Yang, W. Wavelets for Contrast Enhancement of Digital Mammography. *IEEE Eng. Med. Biol. Mag.* **1999**, *14*. [CrossRef]
48. Kaya, B.; Can, Y.B.; Timofte, R. Towards Spectral Estimation from a Single RGB Image in the Wild. *arXiv* **2018**, arXiv:cs.CV/1812.00805].
49. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*. [CrossRef]

electronics

MDPI

Article

An Efficient Hybrid Fuzzy-Clustering Driven 3D-Modeling of Magnetic Resonance Imagery for Enhanced Brain Tumor Diagnosis

Suresh Kanniappan [1], Duraimurugan Samiayya [1], Durai Raj Vincent P M [2], Kathiravan Srinivasan [2], Dushantha Nalin K. Jayakody [3,4,*], Daniel Gutiérrez Reina [5] and Atsushi Inoue [6]

1 Department of Information Technology, St. Joseph's College of Engineering, Chennai, Tamil Nadu 600119, India; sureshk@stjosephs.ac.in (S.K.); duraimurugans@stjosephs.ac.in (D.S.)
2 School of Information Technology and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu 632014, India; pmvincent@vit.ac.in (D.R.V.P.M.); kathiravan.srinivasan@vit.ac.in (K.S.)
3 School of Computer Science and Robotics, National Research Tomsk Polytechnic University, 634050 Tomsk, Russia
4 Centre for Telecommunication Research, Faculty of Engineering, Sri Lanka Technological Campus, Padukka 10500, Sri Lanka
5 Department of Electronic Engineering, University of Seville, 41092 Sevilla, Spain; dgutierrezreina@us.es
6 Information Systems and Business Analytics Department, Eastern Washington University, Spokane, WA 99202, USA; inoueatsushij@gmail.com
* Correspondence: nalin@tpu.ru

Received: 4 January 2020; Accepted: 5 February 2020; Published: 12 March 2020

Abstract: Brain tumor detection and its analysis are essential in medical diagnosis. The proposed work focuses on segmenting abnormality of axial brain MR DICOM slices, as this format holds the advantage of conserving extensive metadata. The axial slices presume the left and right part of the brain is symmetric by a Line of Symmetry (LOS). A semi-automated system is designed to mine normal and abnormal structures from each brain MR slice in a DICOM study. In this work, Fuzzy clustering (FC) is applied to the DICOM slices to extract various clusters for different k. Then, the best-segmented image that has high inter-class rigidity is obtained using the silhouette fitness function. The clustered boundaries of the tissue classes further enhanced by morphological operations. The FC technique is hybridized with the standard image post-processing techniques such as marker controlled watershed segmentation (MCW), region growing (RG), and distance regularized level sets (DRLS). This procedure is implemented on renowned BRATS challenge dataset of different modalities and a clinical dataset containing axial T2 weighted MR images of a patient. The sequential analysis of the slices is performed using the metadata information present in the DICOM header. The validation of the segmentation procedures against the ground truth images authorizes that the segmented objects of DRLS through FC enhanced brain images attain maximum scores of Jaccard and Dice similarity coefficients. The average Jaccard and dice scores for segmenting tumor part for ten patient studies of the BRATS dataset are 0.79 and 0.88, also for the clinical study 0.78 and 0.86, respectively. Finally, 3D visualization and tumor volume estimation are done using accessible DICOM information.

Keywords: MR brain segmentation; fuzzy clustering; object extraction; silhouette analysis; DICOM processing; 3D modeling

1. Introduction

Brain tumor detection is crucial in medical diagnosis as it provides adequate information about anomalies present in the tissues. This information is necessary to understand the prognosis of the disease and also for treatment planning [1]. Magnetic Resonance Imaging (MRI) procedures help to sense the irregularities of human bodies in three dimensions, non-invasively. In particular, various segmentation techniques are applied to MR brain images by radiographers to identify the extent of abnormality present [2,3]. Recently, many Computer-Aided Detection (CAD) methods are employed for brain tumor detection [4–6]. Subsequently, radiologists anticipate that usage of CAD schemes over brain MR images can advance the diagnostic capabilities with their collaborative effects [7,8].

The Digital Imaging and Communications in Medicine (DICOM) standard image format delivers increased diagnostic relevance. DICOM-compliant MR imaging devices adhere to a specific protocol for archiving and communication of digital medical images. DICOM (.dcm) files afford metadata information such as patient study, equipment settings, and image characteristics-modality, size, bit depth, and dimensions. The DICOM header object is organized as a standard series of tags. These tags are categorized as groups such as image pixel, the image plane, MR/CT Image, and patient information [9,10].

The size of this header differs depending on data elements in each group. For, eg: the image plane module contains various vital parameters, which include image position, slice location, and pixel spacing. From these parameters, the spatial relationship between the slices is computed. DICOM facilitates to create private tags that define data elements accessed within the application created. Various imaging modalities stores digital images in DICOM format, which provides better volume of metadata compared to other formats. DICOM provides harmonization through which the patient under study is wholly analyzed and it also compatible with many commercial toolkits.

The patient dataset is inherently acquired by DICOM-compliant devices. Many methods had been proposed by researchers to segment desired features from the digital images. Intensity-based segmentation methods rely on fixing thresholds and are easier to implement [11]. However, due to high-intensity variations of MR images, the methods yield poor performance and lacks in piecewise continuity. Clustering techniques are standard iterative algorithms that is based on the minimization of an objective function. It considers the pixel intensity values for precisely classifying the image pixels. The extraction of cells or tissues based on morphology, clustering algorithms are used extensively. Many algorithms existing in the literature have the objective to yield better segmentation. With the K-means clustering algorithm [12], a large set of structures is distributed into disjoint and homogeneous clusters. Dhanachandra has attempted image segmentation using a hybrid combination of K-means clustering and the Subtractive Clustering Algorithm [13]. Abdel-Maksoud attempted a combined approach of K-means and Fuzzy C-means clustering technique for brain tumor detection [14]. Kim proposed quantization of full/partial (thickness) tear of rotator cuff tendon using Fuzzy C-Means based classification [15]. Dehariya proposed the segmentation of images using Fuzzy K-means clustering [16]. Gasch implemented Fuzzy k-means clustering as an analytical tool for mining biological perceptions from yeast gene-expression data [17]. Even clustering techniques perform faster computation, a wrong choice of k may produce inaccurate results.

Markov random fields (MRFs) benefit more straightforward implementation by encoding spatial data which expresses a set of parameters for specifying tumor voxels [18,19]. This method is very robust for MR images and their performance entirely depends on spatial constraints and hence not suitable for heterogeneous tissue classes. Statistical pattern recognition based methods also known as atlas-based segmentation methods, are effective only for bi-level segmentation. These approaches require healthier brain atlas that is modified significantly to accommodate the tumor part which may lead to poor results. Hybrid methods utilize the advantage of many models which is used in numerous applications by integrating different models within a system to enhance segmentation accuracy. Fuzzy clustering exhibits excellent performance on images containing homogeneous and

heterogeneous tissue classes [20]. However, fuzzy clustering produces better results by choosing the proper selection of the number of clusters '*k*'.

In the literature, to assess the number of clusters, a metric-based method called silhouette score is used. It evaluates the number of clusters based on their proximity. The silhouette score is interpreted as excellent, moderate, weak and bad splits based on cluster selection. Lleti had attempted to optimize the silhouettes using a genetic algorithm in choosing variables for the K-means cluster examination [21]. Muca determined the optimal number of clusters based on the silhouette index for the K-means algorithm [22]. Robust segmentation based on the finest silhouette scores is performed on a set of DICOM slice sequences that assists in the segregation of abnormal portions from the brain tissue.

Numerous approaches have been proposed for the detection of various objects of interest after segmentation is performed. Zeng proposed K-means with a hybrid active contour model to generate an initial segmentation for segmenting thick-vessels in liver images [23]. Koulountzios developed a simple pipeline for segmenting the whole thoracic aorta into contours such as arch, descending, and ascending aorta from MR DICOM files containing thoracic region [24]. Nekooeimehr proposed a method for tracking and segmenting organ contours using k-means clustering with prior information [25]. Wang has attempted contour refinement using an active contour model to segregate candidate cavernoma sections from brain MR slices [26]. An improved performance utilizing local and global image information for contour detection into a hierarchical region tree [27]. Essadike suggested Van der Lugt correlator-based initial contour to assist an active contour model in extracting tumor boundaries [28].

Morphology is a broad set of non-linear operations that process images that rely on shape and texture classification [29,30]. Ali attempted the K-Means Clustering technique for accounting pixel intensities and locations [31]. The author had applied to dilate and erode morphological operations to abstract the tumor part from the brain tumors, which also aided to eliminate small isolated points. Deng employed morphological operators to enhance the extracted ulcer area from ocular staining images [32].

A comparative investigation between the mined region of interest (ROI) and master segmented (Ground Truth) images is carried out with the well-known image similarity measures [33,34]. The Jaccard and Dice coefficients are calculated to validate the segmentation performed on each slice against their corresponding ground truth object. Modeling 3D view of a patient study requires resampling and image interpolation methods [35] to align the abnormal intensities in the spatial domain geometrically.

The key contributions of this work are summarized as follows:

- This research study uses the advantage of fuzzy clustering (for image enhancement) hybridized with Distance Regularized Level Set technique to effectively mine the region of interest form the brain slices.
- In this work, for each brain slice we have utilized the attributes of DICOM standards such as Image position patient, Pixel spacing and Image orientation patient, which is essential for generating the 3D model of brain structures and volumetric analysis.
- For image enhancement in identifying the objects of interest, fuzzy clustering is employed through proper selection of the number of clusters '*k*' validated using the silhouette metric. The appropriate k is chosen based on the silhouette metric among the number of clusters (*k*) ranging from 2 to 9.
- The proposed work is initially tested on the brain MR series of BRATS dataset for anomaly extraction; its segmentation quality is assessed with image quality, similarity and statistical measures. The average dice scores over ten patient studies for tumor segmentation has given promising results. Further, the procedure is also tested on the clinical MR brain series and validated against expert ground truth.

In this work, the proposed tool is implemented using python open-source language. The proposed methodology is described in Section 2. The obtained results and their relevant findings are demonstrated in Section 3. The conclusions and future scope are discussed in Section 4. A brief video describing the proposed method, its key contributions' and results, is provided in the Supplementary Materials.