



Politecnico
di Bari

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING

Ph.D. Program

SSD: ING-INF/06 - ELECTRONIC AND INFORMATION BIOENGINEERING

Final Dissertation

Computational Imaging for Precision Medicine: the Emergence of Radiomics, Pathomics and Deep Learning

by

Nicola Altini

Supervisors:

Prof. Vitoantonio Bevilacqua, Ph.D.

Prof. Francescomaria Marino, Ph.D.

Coordinator of Ph.D. Program:

Prof. Mario Carpentieri, Ph.D.

Course n°35, 01/11/2019 - 31/10/2022



LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il sottoscritto **Nicola Altini** nato a **Martina Franca** il **08/03/1995**

residente a **Bari (BA)** in via **Giacomo Tauro 3/C** e-mail **nicola.altini@poliba.it**

iscritto al 3° anno di Corso di Dottorato di Ricerca in **Ingegneria Elettrica e dell'Informazione** ciclo **XXXV**

ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo:

Computational Imaging for Precision Medicine: the Emergence of Radiomics, Pathomics and Deep Learning

DICHIARA

- 1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
- 2) di essere iscritto al Corso di Dottorato di ricerca in **Ingegneria Elettrica e dell'Informazione** ciclo **XXXV**, corso attivato ai sensi del *"Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari"*, emanato con D.R. n.286 del 01.07.2013;
- 3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
- 4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archiverà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito <http://www.creativecommons.it/Licenze>), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
- 5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviata/di inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
- 6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
- 7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Luogo e data **Bari, 07/12/2022**

Firma

Il sottoscritto, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

CONCEDE

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Luogo e data **Bari, 07/12/2022**

Firma



Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING
Ph.D. Program

SSD: ING-INF/06 - ELECTRONIC AND INFORMATION BIOENGINEERING

Final Dissertation

Computational Imaging for Precision Medicine: the Emergence of Radiomics, Pathomics and Deep Learning

by
Nicola Altini

Referees:

Prof. Mario Cesarelli
Prof. Leandro Pecchia, Ph.D.

Supervisors:

Prof. Vitoantonio Bevilacqua, Ph.D.
Prof. Francescomaria Marino, Ph.D.

*Coordinator of Ph.D. Program:
Prof. Mario Carpentieri, Ph.D.*

Course n°35, 01/11/2019 - 31/10/2022

Abstract

The purpose of this Ph.D. thesis is to illustrate the research works carried out during the conceptualization, design, implementation, and evaluation of novel Clinical Decision Support Systems (CDSSs) based on Radiomics, Pathomics and Deep Learning (DL) techniques.

CDSSs can be effective systems for implementing Precision Medicine into clinical practice since they permit the objective and repeatable evaluation of patients. Precision Medicine can enable the improvement of the healthcare system by employing a personal healthcare process for the health status of an individual patient, which evolves in a unique way.

The methodologies concerning CDSSs were developed with different underlying goals: improvement of the clinical results, availability and usability of the method, and feasibility of the integration into the routine clinical practice. The applications considered span from Radiology to Digital Pathology. Tasks under consideration in Medical Imaging applications, from a computer vision perspective, concerned object detection, instance segmentation, semantic segmentation, color normalization, and characterization and classification of regions of interest. Data under consideration were either provided by local hospitals or obtained from public repositories. Validation of the developed systems has been done in accordance with the physicians. Moreover, the explainability of the realized systems has been investigated, by analyzing features' structure or by means of perceptive saliency maps.

In the aforementioned scenario, the main purpose of this thesis is to develop new systems based on Deep Learning, Radiomics and Pathomics for the processing and analysis of medical images. Computational Imaging is a promising methodology to incorporate in the framework of Precision Medicine. Indeed, it creates the possibility to characterize the lesions in large datasets of images belonging to Radiology and Digital Pathology domains in an effective way, offering a personalized evaluation of the patient.

Merits and shortcomings regarding DL in the field of Medical Imaging have been investigated for applications in Radiology and Digital Pathology. Technical contributions include devising novel algorithms, improving existing workflows, and assembling complex CDSSs by combining in an original and effective way different techniques proposed in the literature.

In the Radiology domain, the following tasks have been tackled for what concerns applications related to Image-guided Surgery (IGS): liver segmentation, including also the classification into anatomical segments; vertebrae segmentation and identification; prostate segmentation and registration for image fusion. Radiomics has been exploited for characterizing lung lesions in COVID-19 patients, in order to discover a prognostic signature for those with a higher risk of developing pulmonary thromboembolism.

With regard to Digital Pathology, applications included colorectal cancer (CRC) tissue classification; hematoxylin and eosin (H&E) stain color normalization; nuclei segmentation and detection; glomeruli lesions classifications according to Oxford score for IgA nephropathy patients. These automatic pipelines for histological data analysis can enable Pathomics, allowing the objective quantification and evaluation of tissue patterns.

The developed solutions in all these scenarios were put in comparison with state-of-the-art approaches proposed in the literature, and were validated with physicians when possible. In many cases, data have also been collected from local institutions.

This thesis work is organized into five chapters. Chapter 1 introduces the objective and the technical contribution of the thesis. Chapter 2 describes the state-of-the-art in all the considered clinical scenarios, with a particular focus on Radiology and Digital Pathology, encompassing emerging trends such as Radiomics and Pathomics. Chapter 3 describes the contributions proposed in the Radiology field. In particular, IGS applications concern liver segmentation and classification into segments, vertebrae segmentation and identification, and prostate segmentation and registration. Also, a Radiomics-based analysis of lung lesions of patients diagnosed with COVID-19 is presented. Chapter 4 presents the contributions proposed in the field of Digital Pathology, concerning tissue segmentation, normalization and classification, and detection of objects of interest, such as nuclei of cells. Lastly, final remarks and considerations for future works are drawn in Chapter 5.

Table of contents

List of figures	ix
List of tables	xii
1 Introduction	1
1.1 Objective and Research Question	3
1.2 Contribution	4
1.3 Thesis Outline	5
2 State of the art	6
2.1 Deep Learning	6
2.1.1 Classification	7
2.1.2 Semantic Segmentation	9
2.1.2.1 Encoder-decoder Architectures	9
2.1.3 Object Detection	10
2.1.4 Generative Adversarial Networks	12
2.1.4.1 Conditional GANs: <i>pix2pix</i>	14
2.1.4.2 Cycle-Consistent GANs: <i>CycleGAN</i>	15
2.1.5 Self-Supervised Learning	16
2.1.6 Conditional Random Fields	18
2.1.7 Quality Measures	19
2.1.7.1 Classification	19
2.1.7.2 Semantic Segmentation	19
2.1.7.3 Object Detection	21
2.1.7.4 Image Similarity	23
2.1.8 Explainable AI Methods	23
2.1.8.1 Perceptive XAI	24

2.1.8.2	Mathematically Explained XAI	25
2.2	Radiomics and Deep Learning in Radiology	26
2.2.1	Datasets and Challenges	29
2.2.1.1	Single-Organ Segmentation	30
2.2.1.2	Multi-Organ Segmentation	32
2.2.2	Segmentation Workflow	33
2.2.3	Liver Parenchyma Segmentation	36
2.2.3.1	Pure CNN approaches	38
2.2.3.2	CNNs combined with post-processing	38
2.2.3.3	Adversarial Loss setups	40
2.2.3.4	Summary	40
2.2.4	Liver Segments Classification for Surgical Planning	41
2.2.5	Vertebrae Segmentation and Identification	44
2.2.5.1	Related Work	45
2.2.6	Prostate Fusion Biopsy	47
2.2.7	Lung COVID-19 Characterization	50
2.2.7.1	Radiomic Features Definitions	52
2.3	Pathomics and Deep Learning in Digital Pathology	55
2.3.1	Stain Color Normalization in CRC Histopathology	56
2.3.1.1	Related Works	58
2.3.2	Nuclei Detection from Histopathological Images	60
2.3.3	Computer-Aided Diagnosis for IgA-GN Oxford Classification	61
2.3.3.1	MEST-C Score	62
2.3.3.2	Related Works	63
3	Radiomics and Applications of Deep Learning in Radiology	64
3.1	Liver Segments Classification for Surgical Planning	64
3.1.1	Materials	65
3.1.2	Workflow	66
3.1.2.1	Focal Loss	67
3.1.2.2	Focal Dice Loss	68
3.1.2.3	Post-Processing	71
3.1.3	Results	73
3.1.4	Summary	76
3.2	Vertebrae Segmentation and Identification	79

3.2.1	Materials	81
3.2.2	Workflow	81
3.2.2.1	Spine Segmentation	81
3.2.2.2	Vertebrae Identification	83
3.2.2.3	Visualization Tool	89
3.2.3	Results	89
3.2.4	Summary	94
3.3	Prostate Fusion Biopsy	94
3.3.1	Materials	95
3.3.2	Workflow	96
3.3.2.1	MRI Segmentation	98
3.3.2.2	TRUS Segmentation	99
3.3.2.3	MRI-TRUS Registration	106
3.3.3	Results	109
3.3.3.1	Segmentation	109
3.3.3.2	Registration	110
3.3.3.3	Discussion	112
3.3.4	Summary	113
3.4	Lung COVID-19 Characterization	114
3.4.1	Materials	115
3.4.2	Workflow	115
3.4.3	Radiomic Features	118
3.4.4	Results	119
3.4.4.1	Segmentation Experimental Results	119
3.4.4.2	Radiomics Lung Characterization	122
3.4.4.3	Discussion	129
3.4.5	Summary	129
4	Pathomics and Applications of Deep Learning in Digital Pathology	131
4.1	Stain Color Normalization in CRC Histopathology	131
4.1.1	Materials and Methods	132
4.1.1.1	Datasets	132
4.1.1.2	Image Processing Normalization	133
4.1.1.3	Deep Learning Normalization	136
4.1.1.4	Stain Normalization Workflow	140

4.1.1.5	Generated Image Tiles Quality Assessment	140
4.1.1.6	Experimental Design	141
4.1.2	Experimental Results	143
4.1.2.1	Stain Color Normalization	143
4.1.2.2	Multi-class Classification	145
4.1.2.3	Relationships between Normalization and Classification .	151
4.1.3	Discussion	157
4.1.4	Summary	158
4.2	Nuclei Detection from Histopathological Images	159
4.2.1	Materials and Methods	161
4.2.1.1	Datasets	161
4.2.1.2	NDG-CAM	162
4.2.1.3	Instance Segmentation	169
4.2.1.4	Implementation Details	169
4.2.1.5	Combined Model	170
4.2.1.6	Evaluation Metrics	170
4.2.2	Results	171
4.2.2.1	Discussion	175
4.2.3	Summary	178
4.3	Computer-Aided Diagnosis for IgA-GN Oxford Classification	179
4.3.1	Materials	180
4.3.1.1	Dataset Composition	180
4.3.2	Methods	183
4.3.2.1	Classical Features	183
4.3.2.2	Convolutional Neural Networks	186
4.3.3	Results	189
4.3.3.1	Classical Features	189
4.3.3.2	Convolutional Neural Networks	189
4.3.4	Summary	195
5	Conclusion	197
	My Publications	201
	References	205

List of figures

2.1	Encoder-decoder architectures: a graphical comparison	11
2.2	GAN architectures: a graphical comparison	13
2.3	Deep Learning Publications in Medical Imaging	27
2.4	Liver, Kidney, and Spleen image examples	34
2.5	Segmentation Workflow: A Schematic Representation	37
2.6	Prostate gland in TRUS image	49
3.1	Medical Segmentation Decathlon Task 08 dataset	65
3.2	V-Net architecture for multi-class segments' classification	66
3.3	<i>FDL</i> plots for the different considered formulations	70
3.4	Liver segments' anatomy	72
3.5	Qualitative results for liver segments' classification	73
3.6	Normalized confusion matrices	77
3.7	Vertebrae segmentation workflow	83
3.8	Vertebrae identification workflow	85
3.9	Workflow of the morphological analysis	85
3.10	Sample <i>ConnectedComponentImageFilter</i> output	86
3.11	Sample outputs of cluster analysis and arch/body coupling	88
3.12	Visualization tool for visualizing CT scans and masks	90
3.13	Examples of binary spine segmentation	91
3.14	Samples of the final output of the centroids' prediction	92
3.15	Sample outputs of meshes of multi-class vertebrae segmentation	92
3.16	Comparison between Payer et al. and the proposed method for the vertebrae labeling task	93
3.17	Samples of images from MRI and TRUS modalities	96
3.18	Workflow for TRUS and MRI segmentation and subsequent image fusion	97
3.19	Deformable superellipse modeling capabilities examples	101

3.20	Workflow for TRUS segmentation with the developed methodology	105
3.21	Image registration pre-processing	107
3.22	Proposed registration workflow	108
3.23	Results for prostate segmentation from MRI	109
3.24	Qualitative results for prostate segmentation from TRUS	111
3.25	Image fusion workflow	112
3.26	CT scans dataset of COVID-19 patients	116
3.27	Workflow for segmentation of lung parenchyma and lesions	120
3.28	Example CT scans and predictions	121
3.29	Radiomic features in the 2D PCA and t-SNE spaces	124
3.30	Cluster map for the radiomics features	127
3.31	Violin plots for the subset of 9 significant and uncorrelated radiomic features	128
4.1	Datasets exploited for the stain normalization and the downstream classifica- tion task	134
4.2	Example of artifacts introduced by classical stain color normalization methods	135
4.3	Workflow employed for the stain color normalization study	144
4.4	Normalization examples with the various UI2IT GANs considered	145
4.5	Boxplots for SSIM between saturation channel of normalized images versus original one	146
4.6	Saturation differences between the considered UI2IT methods	147
4.7	Bar Plots with error bars of PSNR, SSIM, LPIPS for N(SD1) vs SD1 and N(SD2) vs SD2	148
4.8	Bar Plots with error bars of FID for N(SD1) vs SD2, N(SD1) vs N(SD2), N(CDT) vs CDV, and N(CDT) vs N(CDV)	149
4.9	Line Plots of PPQ for SD1 vs N(SD1) and SD2 vs N(SD2)	149
4.10	Accuracy for the various normalization methods on the CDT and CDV datasets	152
4.11	Embedding plot obtained applying t-SNE to features extracted from DenseNet201 on the CDV dataset, with the FastCUT normalization methodology	153
4.12	Correlations between classification metrics and FID	156
4.13	Correlation between FID and PPQ, and between LPIPS and PPQ	156
4.14	Pipeline adopted for training and validation	163
4.15	Sample images of datasets for nuclei detection	164
4.16	Semantic segmentation output for nuclei images	166
4.17	NDG-CAM Detection workflow	168

4.18	Example of calculation of evaluation metrics for object detection	171
4.19	Violin plots for the detection metrics calculated per case	175
4.20	Examples of the NDG-CAM method on the data from the Pathology Department of IRCCS Istituto Tumori “Giovanni Paolo II”	177
4.21	Examples of centroid detection on the validation sets V1 and V4	178
4.22	The glomeruli dataset distribution	181
4.23	IgA Glomerulonephritis — Glomerular patterns	182
4.24	Workflow for preprocessing and feature extraction from images	187
4.25	Transfer Learning Workflow for Glomeruli MESC Classification	188
4.26	Grad-CAM sample results for S-score	193

List of tables

2.1	Binary classification confusion matrix	20
2.2	Single-Organ Datasets Summary	32
2.3	Multi-Organ Datasets Summary	33
2.4	Liver Segmentation with Deep Learning: a comparison	42
2.5	Performance comparison for Liver Segmentation	43
3.1	Class distribution for the selected training set	66
3.2	Segments' inconsistency table	71
3.3	Quantitative results for the models trained with the FDL^A formulation . . .	74
3.4	Quantitative results for the models trained with the FDL^B formulation . . .	74
3.5	Quantitative results for the models trained with the FDL^W formulation . . .	75
3.6	Quantitative results for the models trained with the Dice loss	75
3.7	Literature overview for liver segments' classification	78
3.8	Comparison between the proposed method and related works	80
3.9	Spine segmentation datasets	82
3.10	Spine binary segmentation results	90
3.11	Prostate gland datasets description	96
3.12	Quantitative metrics results for MRI prostate segmentation with nnU-Net .	110
3.13	Quantitative metrics results for TRUS prostate segmentation with the pro- posed superellipse-based approach	110
3.14	Quantitative registration results	111
3.15	COVID-19 CT scans datasets	116
3.16	Experimental results for lung segmentation in COVID-19 patients from CT scans	119
3.17	Statistics for feature distribution between PTE and non-PTE cases	126
3.18	Literature overview for COVID-19 image tasks	126

4.1	Mean and standard deviation of PSNR, SSIM, LPIPS for N(SD1) vs SD1 and N(SD2) vs SD2	146
4.2	FID comparison for the various normalization methods	150
4.3	Mean and standard deviation of PPQ for N(SD1) vs SD1 and N(SD2) vs SD2	151
4.4	Accuracy on the CDT and CDV classification datasets	154
4.5	Classification Mean Precision, Recall, and Dice (mPrecision, mRecall, and mDice, respectively)	155
4.6	Summary of datasets for nuclei	162
4.7	Performance comparison between considered network architectures for semantic segmentation	173
4.8	Comparison of methods for Nuclei Detection	174
4.9	Topologies for the ANN configuration	188
4.10	M-score — ANN Results starting from classical features	190
4.11	E-score — ANN Results starting from classical features	190
4.12	S-score — ANN Results starting from classical features	191
4.13	C-score — ANN Results starting from classical features	191
4.14	M-score — ANN Results starting from CNN feature extraction	192
4.15	E-score — ANN Results starting from CNN feature extraction	193
4.16	S-score — ANN Results starting from CNN feature extraction	194
4.17	C-score — ANN Results starting from CNN feature extraction	195

Chapter 1

Introduction

The term Precision Medicine refers to the possibility of improving the healthcare system by applying an individual healthcare process for the health status of each individual patient, which has a unique evolution. In order to accomplish this task, many scientific disciplines, ranging from drug discovery to imaging to genomics, are exploited to realize a rigorous data-driven decision-making process [1]. König et al. [2] defined Precision Medicine as a cyclic process, in which patient data are exploited to realize models which are relevant from a clinical perspective.

In the last years, technological advancement and the increasing availability of clinical images, signals, and data have led to the possibility of pairing the traditional evaluation made by specialists with Clinical Decision Support Systems (CDSSs). A CDSS is a software system with the purpose to empower the medical decision-making process by exploiting clinical knowledge, health and patient information [3]. CDSSs can be effective systems to bring Precision Medicine into clinical practice, by allowing the objective and repeatable assessment of patients.

According to Shortliffe et al. [4], in order to employ a CDSS in the clinical practice, designers should consider different requirements and capabilities: (i) a CDSS should not work as a black box, the operator needs to have insights on the mechanisms underlying computer decisions; (ii) a CDSS should be efficient from a time complexity perspective, in order to make it feasible the adoption in busy clinical environments; (iii) a CDSS should be simple to use, without requiring too much time for specialized training; (iv) a CDSS should solve problems which are relevant to the real-world clinical practice; (v) a CDSS should not replace the clinicians, but aid them in their decision-making process, by providing objective evaluations; (vi) a CDSS should be sound from a scientific perspective, therefore

reproducibility, usability, safety, and reliability must be enforced through rigorous scientific research.

Medical Imaging is taking an increasingly important role in the medical practice since it can improve the processes involved in the diagnosis, treatment, and follow-up of patients. The effectiveness of Medical Image Analysis has a pivotal role in the integration of Precision Medicine into clinical practice. Indeed, Medical Imaging can help to portray not only the anatomy of the patient but also the physiology and metabolism, thus enabling patient-specific procedures and therapies [5].

Computational Imaging has a long history for medical applications, as cell biology [6]. However, only with emerging paradigms, starting from the last decade, imaging has become an effective tool to realize advanced CDSSs, that can really enable Precision Medicine. The most important trends for CDSSs based on Computational Imaging include Deep Learning (DL) [7], Radiomics [8] and Pathomics [9].

DL is very appreciated in this context, especially because it offers the possibility to learn hierarchical feature representations only from the data, instead of making an expensive design of handcrafted features based on domain-specific knowledge. Radiomics and Pathomics refer to the processing of a large number of quantitative features (either handcrafted or deep) extracted from medical images and are revolutionizing the radiologists' and pathologists' workflows, respectively.

Tumors and lesions show strong phenotypic differences which can be visualized in a non-invasive way through Medical Imaging techniques. Radiomics is an increasingly growing paradigm that consists of the extraction of a large number of quantitative features from images with the purpose to obtain a global quantification of phenotypes of lesioned regions, allowing to unveil and quantitatively describe tumor patterns and characteristics [10–12]. Radiomic features give information about gray-level patterns and inter-pixel relationships. Moreover, spectral and shape properties can be extracted from the same regions of interest. These features can be further exploited to realize computational models based on advanced Machine Learning techniques so that they can be used for diagnosis, prognosis, and treatment guide [8, 13, 14].

In the context of histopathological Whole Slide Images (WSIs) examination, Pathomics is surging in the last years as a powerful tool to unveil characteristics from tissue samples. Pathomics refer to the data which can be extracted from image analyses to obtain quantitative features that can characterize phenotypic features of tissue samples in WSIs [9]. Pipelines for WSI image analysis involve several steps, usually comprising detection, segmentation, and classification. CDSSs in this domain can complement the classical histopathological

examination carried out by expert pathologists with quantitative features that are correlated with different phenotypic characteristics [9].

Deep Learning belongs to a broader family of Machine Learning methods based on Artificial Neural Networks (ANNs). DL involves the adoption of computational models with multiple layers which can realize hierarchical feature extraction and processing from raw data, allowing to obtain state-of-the-art results in such fields where it is challenging to design handcrafted algorithms, such as visual object recognition and speech recognition [15].

The awareness of the opportunities granted by a wise adoption of Convolutional Neural Networks (CNNs) might have pivotal importance in diagnostic radiology, allowing incrementing radiologists' performances and improving patients' care [16]. Small sample sizes can have a detrimental effect on CNN models, but generative models, and in particular Generative Adversarial Networks (GANs), can be exploited to mitigate this issue [17].

The exploitation of hierarchical data representations learned with Deep Learning models applied to histopathological images might aid in accomplishing tasks such as segmentation or detection of objects of interest [18, 19], tissue classification [20], tissue color normalization [21], nuclei segmentation and detection [22, 23], predictions of genomic patterns [24] and survival analysis [25].

1.1 Objective and Research Question

In the last decade, Deep Learning techniques have revolutionized the field of computer vision, starting from the milestone work of Krizhevsky et al. [26]. Soon, the biomedical engineering field benefited from such approaches, which allowed the implementation of accurate and fast systems for Clinical Decision Support [7], Computer-Aided Diagnosis (CAD), and Image-guided Surgery (IGS) [27] relying on CNNs models [16], Transformers [28], or GANs [29].

Despite the dramatic improvements carried out by these new models in the CDSSs field, there are still questions to be addressed. Testing in the clinical practice needs to be extended, in order to tackle problems linked to data heterogeneity. Indeed, data coming from different hospitals or laboratories can have inherent differences that can lead to biased models [30]. Furthermore, those systems are often considered black-boxes, which can hinder their deployment in the healthcare system. To create a trust for the physicians so that they can use CDSSs in the routine practice, Explainable Artificial Intelligence (XAI) techniques come in handy [31]. Moreover, also patients have the right to know the dynamics underlying evaluations performed by Deep Learning-based CDSS. For what concerns legislation, the

EU General Data Protection Regulation (GDPR), with Articles 13 and 14, decree that when data subjects are profiled, and this concept also applies to the medical field, they have a right to “meaningful information about the logic involved” [32].

Based on the prior discussion concerning opportunities and difficulties of Deep Learning-based systems for clinical practice, the main objective of the research conducted for this thesis has regarded the conceptualization, development, and implementation of CDSS workflows to advance the state-of-the-art in this field. The pipelines were developed with different underlying goals: improvement of the clinical results, availability and usability of the method, and feasibility of the integration into the routine clinical practice. Moreover, the systems have been made simple to use, and when possible, explainability has been investigated.

The applications considered span from Radiology to Digital Pathology. Tasks under consideration in Medical Imaging applications, from a computer vision perspective, concerned object detection, instance segmentation, semantic segmentation, color normalization, and characterization and classification of regions of interest. Data under consideration were either provided by local hospitals or obtained from public repositories. Validation of the developed systems has been done in accordance with the physicians.

1.2 Contribution

In the aforementioned scenario, the main purpose of this thesis is to develop new systems based on Deep Learning, Radiomics and Pathomics for the processing and analysis of medical images [16, 22, 33, 34]. Merits and shortcomings regarding these techniques in the field of Medical Imaging have been investigated for applications concerning Radiology and Digital Pathology.

Technical contributions include the development of novel algorithms to tackle clinically relevant tasks, the improvement of existing pipelines and methodologies, the statistical analysis of newly collected medical data, and the construction of CDSSs that originally and effectively combine different existing techniques. Overall, the CDSSs and Computational Imaging-based systems developed throughout this thesis have applications that span a wide range of clinical scenarios.

In the Radiology domain, the following tasks have been tackled for what concerns applications related to IGS: liver segmentation, including also the classification into anatomical segments [35]; vertebrae segmentation and identification [36]; prostate segmentation and registration for image fusion [33]. Radiomics has been exploited for characterizing lung

lesions in COVID-19 patients, in order to discover a prognostic signature for those with a higher risk of developing pulmonary thromboembolism [37].

Concerning Digital Pathology, applications covered colorectal cancer (CRC) tissue classification [20]; hematoxylin and eosin (H&E) stain color normalization; nuclei segmentation and detection [22]; glomeruli lesions classifications.

The developed solutions in all these scenarios were put in comparison with the state-of-the-art approaches proposed in the literature and validated with physicians when possible. Also, XAI techniques have been considered to understand the mechanisms underlying the developed systems, either by analyzing the features' structure or with perceptive saliency maps.

1.3 Thesis Outline

This thesis work is organized into five chapters. This chapter introduces the objective and the technical contribution of the thesis. Chapter 2 describes the state-of-the-art in all considered clinical scenarios, with a particular focus on Radiology and Digital Pathology, encompassing emerging trends such as Radiomics and Pathomics. A significant consideration is also reserved for Deep Learning methodologies. Chapter 3 describes the contributions proposed in the Radiology field, including liver segmentation and classification into segments, vertebrae segmentation and identification, prostate segmentation and registration for IGS applications, and the Radiomics approach for COVID-19 lung lesions characterization. Chapter 4 presents the contributions proposed in the field of Digital Pathology, concerning tissue segmentation, normalization, and classification. Lastly, final remarks and considerations for future works are drawn in Chapter 5.

Chapter 2

State of the art

2.1 Deep Learning

Inside the realm of Machine Learning methods founded on ANNs, primary importance is given to Deep Learning methods. These techniques involve the exploitation of computational models that, through multiple levels of abstraction, can be used for the processing of raw data. In fields such as computer vision and speech recognition, where it is infeasible or difficult to design handcrafted features, the potentialities of Deep Learning, which allows the learning of hierarchical features in an end-to-end process, are particularly appreciable [15].

The most well-known Deep Learning architectures, as Convolutional Neural Networks [38], Autoencoders, Deep Belief Networks and Restricted Boltzmann Machines, have been recognized advantageous in a broad spectrum of applications of big data analysis: pattern recognition, natural language processing, computer vision, and recommender systems [39, 40].

Starting from the last decade, a colossal quantity of literature have developed concerning the design and implementation of classifiers based on DL for natural image processing [26, 41–48]. Even though these efforts are very appreciable and have led to a substantial progress in the imaging field, a lot of work has still to be carried out for generalizing these models for medical imaging applications.

In recent years, DL techniques are becoming predominant in the whole medical image analysis field, for tasks that comprises of registration, image generation, image enhancement, retrieval, detection, segmentation, and classification. Applications concern almost all areas in the medical domain: neuro, ophthalmic, breast, cardiac, abdominal, musculoskeletal among others [49].

In the diagnostic radiology domain, a wise exploitation of proper DL models, as CNNs, can lead to an improvement in the radiologists' performances and patients' care. Sadly, most models developed in this context suffer from issues as small training and validation cohorts, likely leading to overfitting or biased models [17].

For what concerns systems aimed at improving pathologists' workflow, the main issues to take into consideration include differences in fixation, staining, and sectioning of the whole slide images [50]. Moreover, only a small number of labeled images is available for developing CDSS in the Digital Pathology domain. Commonly used DL architectures require pixel-level (semantic segmentation tasks) or patch-level (classification tasks) annotations, but in the majority of cases only WSI-level annotations are available [51]. Only expert pathologists can provide accurate annotations for these images, thus further increasing time and costs for the development [52].

In the following of this section, Deep Learning methodologies, targeted at the development of CDSSs enabling Precision Medicine, are presented and analyzed. Section 2.1.1 details CNN models employed for image classification tasks. Section 2.1.2 introduces architectures devoted into semantic segmentation, that is the task of performing a classification at pixel-wise or voxel-wise level. A special focus is given to encoder-decoder architectures, since they achieved a large success in medical applications. Section 2.1.3 briefly discusses CNN architectures which can be used for object detection and instance segmentation. Section 2.1.4 explicates Generative Adversarial Network setups, that, for instance, can be exploited for image segmentation and data augmentation. Section 2.1.5 summarizes the emerging paradigm of Self-Supervised Learning (SSL), which is a powerful methodology for learning representations useful for the handling of downstream tasks starting by unlabeled data. Section 2.1.6 surveys Conditional Random Fields (CRFs) and their relationship with DL, since they are often exploited for post-processing stages of semantic segmentation architectures. Section 2.1.7 presents the main quality measures which are used in the assessment of detection, segmentation, classification and normalization tasks. Section 2.1.8 presents explainability and interpretability techniques which come in handy for understanding how deep models work under the hood.

2.1.1 Classification

CNNs have a pivotal importance in several tasks related to pattern recognition. A lot of literature has developed around the problem of realizing effective CNN image classifiers [26, 41–43, 46, 48, 53, 54].

In the realm of histopathological imaging, applications encompass segmentation and detection of glomeruli from kidney biopsies [18, 19], breast biopsy H&E images classification [55], and CRC tissue multi-class classification [20]. For what concerns processing of radiographic images, applications include liver vessels segmentation [56], liver segments classification [35], and COVID-19 lung and lesions segmentation [37].

Transfer learning allows to mitigate the huge need for data required when training deep CNNs. A comprehensive survey of the transfer learning paradigm, which consists in exploiting data that can be easier collected than the data of the problem under investigation, can be found in [57], whereas one concerning deep transfer learning is [58]. In histopathology, there is a tendency in adopting models pre-trained on ImageNet as feature extractors, as reported by numerous studies in literature [20, 24, 59–63]. Although histopathological tiles look dissimilar from everyday life pictures, there are common basic structures that are shared, i.e. arcs and edges [52]. Since the first layers of CNNs seize this kind of elementary patterns, deep transfer learning can be exploited also for Digital Pathology images.

Commonly used CNN models to perform feature extraction or end-to-end classification are described in the remainder of this Section.

GoogleNet, also known as **Inception-v1**, proposed by Szegedy et al. [43], introduced the inception module, that contains convolutional blocks which are executed in parallel; subsequent versions include **Inception-v2** and **Inception-v3** [64], which are iterative improvements over, and exploit techniques as factorization methods for convolutions and tricks for the exploitation of auxiliary classifiers.

ResNet, proposed by He et al. [54], introduced a deep CNN with 152 layers. The stem idea lying inside is the adoption of skip connections, i. e. the signal entering into a layer is also summed to the output of a higher-level layer. This enables residual learning, that is modelling $f(x) = h(x) - x$ rather than simply the target function $h(x)$, thus obtaining a huge speed gain in the training phase.

DenseNet, proposed by Huang et al. [65], is a model in which, for every layer, the feature maps of all preceding layers are used as input into the subsequent layers using the concatenation operation.

SqueezeNet, proposed by Iandola et al. [66], is a model used in systems that have low computational power. The low dimensionality of this model can result in lower accuracy, which is not desirable in medical tasks, but it also leads to reducing the risk of overfitting and hence a better generalization capability in scenarios with low data available.

2.1.2 Semantic Segmentation

Semantic segmentation refers to the process of labeling every pixel of an input image in semantic meaningful classes, without distinguishing between different object instances. Numerous challenges have to be tackled for solving this task, but a large number of applications can benefit from such a paradigm. A not exhaustive list comprises of human-computer interaction, robotics, agriculture, medical research, and autonomous vehicles [67]. Surveys with definitions, methodologies, techniques, datasets, and applications have been realized in literature [67–69].

The semantic segmentation framework is equivalent to solve an image-to-image conversion problem, where the input image is the source image, and the output image is a map in which every pixel has an intensity which is associated to a meaningful semantic class. In order to accomplish this conversion process, most architectures have an encoder-decoder structure. Feature extraction is carried out by the encoder, reducing the spatial dimensions while enlarging the depth (i.e., the number of feature maps). On the other side, the goal of retrieving the spatial information from the output of the encoder is given to the decoder. Popular architectures embracing the encoder-decoder paradigm include SegNet [70], V-Net [71], U-Net [72], U-Net 3D [73], and Y-Net [74].

Long et al. [75] were the first to introduce the idea of directly adopting CNN for semantic segmentation tasks, with the Fully Convolutional Network (FCN). In their milestone paper, they proved that existing state-of-the-art in the context of semantic segmentation was outperformed by CNNs trained end-to-end for pixelwise classification. The idea behind FCN is that of altering the backbone architecture of existing CNNs (like those employed for classification tasks), so that they can process input of variable size and return consistently-sized output, assuring an efficient mechanism for learning and inference.

The realm of architectures for semantic segmentation is not limited to encoder-decoder structures. Novel ideas include Atrous convolution [76] and Atrous Spatial Pooling [77], as implemented in the latest versions of DeepLab models [76, 78–80].

2.1.2.1 Encoder-decoder Architectures

Badrinarayanan et al. developed a semantic segmentation model founded on the encoder-decoder paradigm [70]. Convolutional and max-pooling layers are the building blocks of the encoding stage, taking advantage of 13 layers from the VGG-16 [41]. The upsampling process of SegNet, carried out in the decoding stage, exploits the max-pooling indices stored from the encoder. This procedure is referred to as max-unpooling. Convolutional layers of

the decoder are placed symmetrically with respect to those of the encoder. In order to predict the semantic output class, a K-class softmax classifier is placed as last layer.

Ronneberger et al. proposed the U-Net, a popular semantic segmentation network tailored at biomedical image segmentation, at MICCAI (Medical Imaging Computing and Computer Assisted Intervention) conference in 2015 [72]. In analogy to the SegNet model, U-Net employs an encoder-decoder paradigm. Instead of implementing max-unpooling, the decoder path takes advantage of transposed convolution layers, that allow a learning process for the upsampling of feature maps. A fundamental improvement offered by the U-Net architecture is the introduction of skip connection between the layers of the encoder and the decoder, enabling to retain the information concerning localization during the whole learning process. At the end, feature maps' depth is converted to the opportune number of classes through a 1×1 convolution layer.

Cicek et al. proposed a natural extension of the U-Net architecture to 3D medical image processing [73]. Given its success and popularity, an enormous number of variants of the base U-Net architecture have been proposed in the last years [71, 73, 74, 81–88]. Literature reviews have become necessary to keep track of the state-of-the-art concerning U-shaped architectures [89].

Metha et al. introduced the Y-Net model, which can be thought of as a generalization of the U-Net architecture since it adds a parallel branch that enables discriminative map generation. Considering that it also supports convolutional block modularity, the user can tune the efficiency of the method without architectural modifications [74].

Milletari et al. conceived V-Net, another popular variation of U-Net, tailored at 3D medical image segmentation tasks [71]. This model replaces the max-pooling layers exploited during downsampling with down-convolutions, namely convolutions with stride $2 \times 2 \times 2$ and kernel-size $2 \times 2 \times 2$. Moreover, PReLU [45] activations are exploited for introducing non-linearities, and residual connections are employed also inside the encoding and the decoding paths. Batch-Normalization (BN) layers after convolutional layers have been added in later works to ease the learning process [86].

Figure 2.1 depicts a graphical comparison between the peculiarities of the considered encoder-decoder models.

2.1.3 Object Detection

Object detection is a computer vision task which consists in determining all the the bounding boxes for the objects of interest which appear in the image. Instance segmentation further

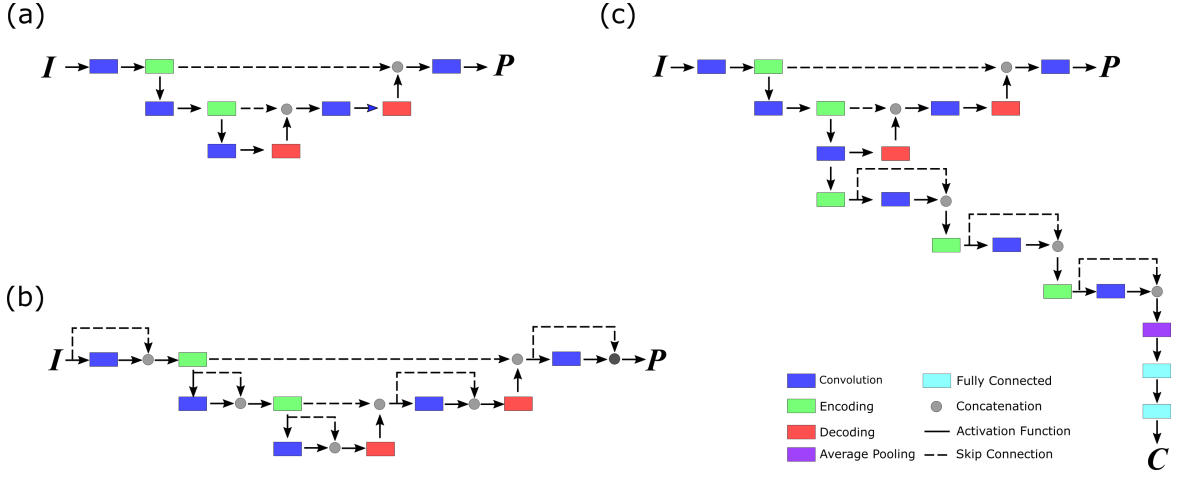


Fig. 2.1 *Encoder-decoder architectures: a graphical comparison.* (a) U-Net-like architectures: note the presence of skip connections across encoding and decoding paths. (b) V-Net-like architectures: notice the occurrence of skip connections also inside compression and expansion paths. (c) Y-Net-like architectures: acknowledge the existence of an additional compression path for the classifying instances. I denotes the original image to be processed, P the predicted segmentation mask, C the (classification) output of the bottom branch of Y-Net.

extends this task, enabling to retrieve refined segmentation masks for the objects under consideration.

Region-Based Convolutional Neural Network (R-CNN) denotes a family of architectures designed for two-stage object detection. Girshick et al. conceived the original model in 2014 [90], which combines region proposals and CNNs with the aim of achieving object detection. In the first stage, R-CNN generates category-agnostic region proposals which are candidates for containing objects. The regions are then given as input to the CNN which performs the feature extraction for each region. Lastly, the feature vector is fed to a collection of class-specific linear support vector machines (SVMs).

R. Girshick proposed Fast R-CNN, an improvement over the base R-CNN, in 2015 [91]. The entire input image is given as input to the CNN in order to obtain the convolutional feature maps. Region proposals are retrieved from the feature maps, and warped into squares. Reshaping the proposals to a fixed size is managed through an ROI-pooling layer, so that they can be delivered to fully connected layers.

Ren et al. further improved the R-CNN model in 2016, with the design of the Faster R-CNN architecture [92]. Feature maps computed in the first stage of Fast R-CNN can be exploited to spawn region proposals, avoiding to use slower algorithms as selective search, which have no learning capabilities. Faster R-CNN adds a Region Proposal Network (RPN)

after the Fast R-CNN stage of feature maps extraction. The concept behind RPN is to replace pyramids of images or pyramids or filters with anchor boxes. Architecture designers can tune the scale and aspect ratio of anchor boxes according to the expected size of the objects to detect.

He et al. extended the Faster R-CNN model to instance segmentation task, with the development of the Mask R-CNN architecture in 2017 [93]. Mask R-CNN architecture is structured in two main blocks: the backbone architecture, which is involved into feature extraction, and the head architecture, which has the aim of performing bounding box regression, mask prediction, and classification.

2.1.4 Generative Adversarial Networks

Goodfellow et al. introduced GANs in a milestone paper in 2014 [94]. The fundamental idea underlying GANs is to jointly train two models: a generator G , which has the task to understand the data distribution, and a discriminator D , which is devoted in evaluating the probability of a sample belonging to the true training data instead that being a fake generated by G . The generator is trained by maximizing the probability of D making a misclassification. This system is analogous to a two players minimax game. Backpropagation can be exploited for the training process when the discriminator and the generator are ANNs. The optimal solution for this problem consists in G recalling data from the training distribution and D which always outputs $\frac{1}{2}$.

After their introduction, GANs have permeated many computer vision domains, including Medical Imaging, where they have a pivotal role in image reconstruction, cross-modality synthesis, detection, semantic segmentation [95], and classification [29].

Other important applications of GANs in medical image analysis tasks include data augmentation and normalization. GAN-based data augmentation can mitigate unbalanced datasets or avoid problem when dealing with sensitive information. Adopting artificial training data does not lead to decrease in accuracy [96, 97]. Three fundamental GAN frameworks, namely original GANs, pix2pix [98], and CycleGAN [99], are graphically compared in Figure 2.2.

Following paragraphs deepen pix2pix and CycleGAN architectures. Additional information concerning GAN architectures, training procedures, and applications can be found in dedicated reviews [29, 100–102].

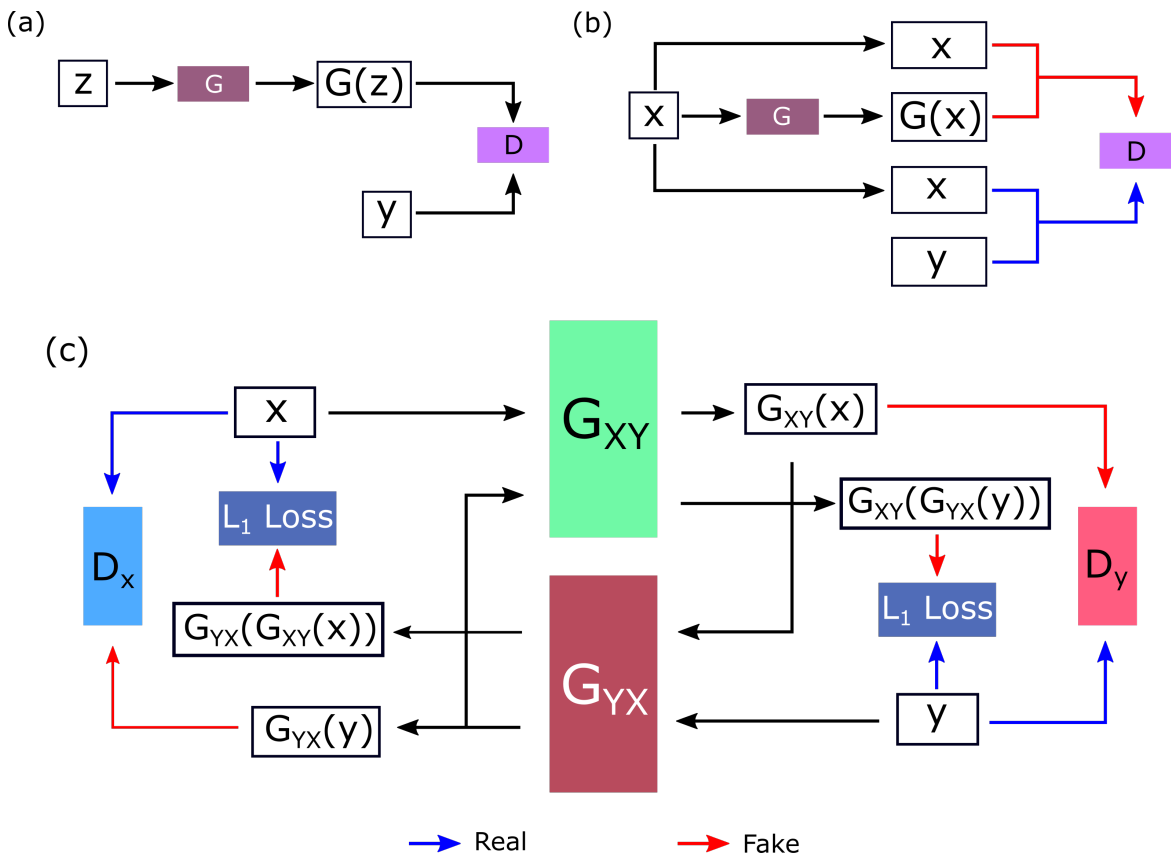


Fig. 2.2 GAN architectures: a graphical comparison. (a) Base GAN: the generator G learns to produce synthetic examples $G(z)$ starting from random noise z , whereas the discriminator D learns to distinguish fake examples, $G(z)$, from real samples, y . (b) cGAN (pix2pix): the generator G learns to produce synthetic examples $G(x)$ starting from conditional examples x , whereas the discriminator D learns to discriminate between fake $x, G(x)$ and real x, y pairs. (c) CycleGAN: there are two generators, G_{XY} and G_{YX} , which maps from X and Y domains, ensuring cycle consistency by transforming the same example forward and backward between the two domains. There are also two discriminators, D_X and D_Y , aimed at recognizing real examples, x and y , respectively, from fake examples, $G_{YX}(y)$ and $G_{XY}(x)$, respectively.

2.1.4.1 Conditional GANs: *pix2pix*

Isola et al. addressed the problem of employing GANs for performing image-to-image translation [98]. The idea of their framework, *pix2pix*, is that many image processing problems which involve the mapping of pixels to pixels, historically handled with dissimilar approaches (i.e., different methodologies, algorithms, and techniques), can be addressed by a unified framework that exploits a conditional adversarial network. The advantage of this approach is that it does not simply learn a mapping from an input image to an output image, but is also able to learn a loss function to train this mapping.

Yang et al. developed a cross-modality image generation framework devoted into synthesis of medical images [103], generalizing over previously proposed approaches, as generating CT images from MRI [104] and mapping low-dose CT into routine-dose CT images [105].

Zhang et al. described an approach for implementing an automatic system for manufacturing plausible colorizations [106]. An important point made by the authors is that the naïve adoption of Euclidean distance between predicted and ground truth pixels may tend to produce a nonrealistic result, since output images tend to be blurred. The key idea of implementing a discriminator is that it can avoid those kind of results since they would seem obvious fakes.

Base GANs are generative models that learn a transformation from a random noise vector z to an output image y , $G : z \rightarrow y$ [94]. Conditional GANs (cGANs), on the other hand, learn a conversion from an observed image x and a random noise vector z , to y , $G : x, z \rightarrow y$. The generator G is trained to construct outputs that cannot be differentiated from “real” images by an adversarially trained discriminator, D , which is trained to do as well as possible at detecting the generator’s “fakes”. An example of the training procedure is depicted in Figure 2.2. The objective of a cGAN can be written as in Equation (2.1):

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log(D(x, y))] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \quad (2.1)$$

where the generator G tries to minimize this objective against an adversarial D that tries to maximize it, i.e. $G^* = \operatorname{argmin}_G(\max_D(L_{cGAN}(G, D)))$. In the original paper, the authors also performed further tests to demonstrate the importance of conditioning the discriminator [98]. A last important observation is that authors decided not to embed directly z in the generator, since it was simply ignored by the generator, as also stated by Mathieu et al. [107]; instead, they modeled the noise with dropout [108, 109].

2.1.4.2 Cycle-Consistent GANs: *CycleGAN*

Cycle-Consistent Adversarial Networks (also known as *CycleGAN*) was proposed by a research group from UC Berkeley [99], with the purpose to tackle the unpaired image-to-image translation problem. With respect to the paired image-to-image translation framework, the unpaired one can be exploited also for domains for which there is no availability of paired training data. The *CycleGAN* architecture learns a mapping $G_{XY} : X \rightarrow Y$ so that the distribution $G_{XY}(X)$ is as similar as possible to the distribution Y according to an adversarial loss. The source and target image domains are denoted as X and Y , respectively. The main issue with a naïve implementation of such approach would be that, for the learning process, the mapping G is tremendously under-constrained. The solution proposed by Zhu et al. is that of coupling the mapping $G_{XY} : X \rightarrow Y$ with an inverse mapping $G_{YX} : Y \rightarrow X$. Hence, the name cycle-consistent, which comes from the cycle consistency loss implemented by enforcing both the mappings at the same time, $G_{YX}(G_{XY}(X)) \approx X$ and $G_{XY}(G_{YX}(Y)) \approx Y$.

The power of this framework has been proven for several applications by Zhu et al., including collection style transfer, season transfer, photo enhancement, object transfiguration, along with others [99]. In the Medical Imaging, examples of applications include noise removal from multiphase coronary CT angiography [110] and stain color normalization for Digital Pathology workflows [21].

In a mathematical way, the aim of *CycleGAN* may be formalized as follows. Given N training samples for domain X , $\{x_i\}_{i=1}^N$, where $x_i \in X$, and M training samples for domain Y , $\{y_j\}_{j=1}^M$, where $y_j \in Y$, the objective is to learn mapping functions between the two domains X and Y . Data distribution can be denoted as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. The *CycleGAN* framework comprehends two mappings $G_{XY} : X \rightarrow Y$ and $G_{YX} : Y \rightarrow X$. Furthermore, there is a couple of adversarial discriminators D_X and D_Y . The aim of D_X is to discriminate between images x and translated images $G_{YX}(y)$. On the other hand, the aim of D_Y is to distinguish between y and $G_{XY}(x)$. The objective function comprises of two types of terms: (i) adversarial losses for matching the distribution of generated images to the data distribution in the target domain; (ii) cycle consistency losses to prevent the learned mappings G_{XY} and G_{YX} from contradicting each other.

The objective adversarial loss for a mapping $G_{XY} : X \rightarrow Y$ and a discriminator D_Y can be expressed as in Equation (2.2).

$$L_{GAN}(G_{XY}, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)} [\log(D_Y(y))] + \mathbb{E}_{x \sim p_{data}(x)} [\log(1 - D_Y(G_{XY}(x)))] \quad (2.2)$$

Hence, cycle consistency loss can be expressed as in Equation (2.3).

$$L_{cyc}(G_{XY}, G_{YX}) = \mathbb{E}_{x \sim p_{data(x)}} [\|G_{YX}(G_{XY}(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data(y)}} [\|G_{XY}(G_{YX}(y)) - y\|_1] \quad (2.3)$$

Finally, the complete objective for *CycleGAN* can be formulated as shown in Equation (2.4):

$$L(G_{XY}, G_{YX}, D_X, D_Y) = L_{GAN}(G_{XY}, D_Y, X, Y) + L_{GAN}(G_{YX}, D_X, Y, X) + \lambda L_{cyc}(G_{XY}, G_{YX}) \quad (2.4)$$

where the parameter λ can be used to address the relative importance of both loss terms. The minimization problem to solve can be expressed as reported in Equation (2.5):

$$G_{XY}^*, G_{YX}^* = \arg \min_{G_{XY}, G_{YX}} \{ \max_{D_X, D_Y} L(G_{XY}, G_{YX}, D_X, D_Y) \} \quad (2.5)$$

The original authors proved that the adoption of both adversarial loss L_{GAN} and cycle consistency loss L_{cyc} is essential for obtaining high-quality results [99].

2.1.5 Self-Supervised Learning

Self-Supervised Learning considers pretext tasks, that can be handled without manual labeling, with the aim to learn convenient data representations, that can be subsequently exploited for solving downstream tasks, which are the ones useful for the final users.

In recent years, SSL is attracting a lot of attention since it can be a valid and successful alternative compared to transfer learning from large datasets. It enables supervised feature learning in scenarios where the supervised problems are self-generated from the unlabeled data available. CNN and other DL architectures can benefit from such a pre-training, with huge performance improvements in cases where there is a scarcity of available labeled data [111].

Furthermore, SSL models have more generalization capability and are more robust to attacks than traditional deep supervised models. SSL advantages may be unnoticed when considering only the overall performances of the CNN on clean data [112]. A comprehensive review encompassing existing empirical methods according to their objective, in the context of SSL, has been performed by Liu et al., with a taxonomy that comprehends three main branches: generative, contrastive, and adversarial [113].

The following of this section presents some applications of SSL in Medical Imaging.

Dhere et al. [114] applied SSL to Kidney Segmentation. A Siamese Convolutional Neural Network (SCNN) is adopted to detect if a given pair of kidney sections from CT scans belongs to the same or to different sides, thus leveraging the anatomical asymmetry of kidneys to obtain a proxy task that can enable effective feature representation learning for the Kidney Segmentation purpose. After the SSL training, the gained information is retained by using one branch of the SCNN as the encoder for the segmentation architecture. With respect to a CNN trained from scratch, this procedure enables an increase in performance and a fast convergence.

Taleb et al. [115] introduced the multimodal puzzle task, in order to realize an SSL method that can exploit multiple image modalities, thus easing the possibility to perform a rich semantic representation learning in this context. The multimodal puzzle, according to the experimental results provided by the authors, can yield an improvement in the semantic representations, if compared to considering each modality independently. These learned representations can be later exploited for performing fine-tuning on different downstream tasks, as Brain Tumor Segmentation and Survival Analysis from MRI (four modalities), Prostate Segmentation from MRI (two modalities), Liver Segmentation from unregistered CT and MRI modalities.

Chung et al. [116] proposed an FCN for Liver Segmentation from CT images. An adaptive self-supervision scheme to derive the essential contour has been exploited to deeply supervise the CNN in order to improve the delineation result.

Chen et al. [111] proposed a novel SSL strategy based on context restoration in order to better exploit unlabeled images. The context restoration strategy allows learning semantic image features which are useful for subsequent tasks of image analysis. The proposed strategy can be simply and effectively implemented and has been validated in three kinds of Medical Imaging tasks: localization, segmentation, and classification.

Stacke et al. [117] investigated the possibility to extend SimCLR [118] to learn useful data representations for histopathology scenarios. The authors considered H&E stained images, and compare how the dissimilarities with the ImageNet dataset affects the SSL objective. The authors state that heuristics which have proven useful for natural images cannot necessarily be extended to histopathology data applications.

2.1.6 Conditional Random Fields

CRF is a probabilistic framework adopted for segmenting and labeling sequence data [119]. With respect to approaches like Hidden Markov Models (HMMs) and Stochastic Grammars, CRFs have important advantages, such as the ability to relax the strong assumptions about independence made by those models.

Krähenbühl et al. proposed a highly efficient approximate inference for fully connected CRF models in which the pairwise edge potentials are defined by a linear combination of Gaussian kernels [120]. According to the results they obtained, dense connectivity at the pixel level dramatically improved segmentation and labeling accuracy. After the rise of CNNs for semantic segmentation (see Section 2.1.2), the most efficient models traditionally combine the structured modeling capabilities of CRFs with the feature extraction power of CNNs.

However, it has to be noted that recently CRF post-processing is being replaced with more fresh approaches [79, 121, 122]. For this reason, Teichmann and Cipolla proposed to add the assumption of conditional independence to the framework of fully connected CRFs [123]. As a result, the authors were able to reformulate the inference in terms of convolutions, so that efficient GPU implementations can be exploited. Moreover, all parameters of convolutional CRFs can be easily optimized using backpropagation. The possibility to employ CRF on top of FCN for improving segmentation quality in 3D multi-modal medical images has been investigated by Monteiro et al. [124]. According to the authors, CRFs are powerful for semantic segmentation of 2D RGB images, but the technique transfers poorly from natural images to medical images.

The core idea behind the approach of combining FCN and CRF is that the latter takes as input the segmentation produced by the CNN as well as the original input images; CRF looks at every possible pair of pixels in the image, differently from how local convolutional filters work, also known as a clique. The graphical model of CRF allows defining every clique not only by spatial distance between pixels but also by their distance in the color space. So, CRF is able to produce a segmentation with much sharper edges compared to those obtained exploiting FCN only [124]. One problem with 3D images is that the third dimension introduces exponentially more pairs of hyper-voxels in the graph. For instance, a 2D square image with a width of N has N^2 cliques in the graph, whereas a 3D image with the same width will have N^3 cliques.

Another problem of CRF concerns the separate training process from the FCN. A possible solution has been considered by [125], where the authors proposed to write the CRF mean-field approximation described in [120] as a Recurrent Neural Network (RNN), which can

be placed on top of the CNN so that the system can be trained end-to-end. Monteiro et al. tried to replicate an analogous concept for 3D medical image segmentation, but the obtained results were not statistically significant [124]. The authors' reasons to explain these issues include the fact that natural images tend to have much higher contrast and much sharper edges than MR images, and edges between objects in natural images tend to be much more well-defined than in MR images.

2.1.7 Quality Measures

2.1.7.1 Classification

The standard method to assess the performances of a statistical classifier is to evaluate the confusion matrix result of the model. For a binary classification problem, it can be seen in Table 2.1. Starting from TP, TN, FP, FN (True Positives, True Negatives, False Positives and False Negatives) introduced in the confusion matrix, it is possible to define the *Accuracy*, as in Equation (2.6) the *Recall* or *Sensitivity*, as in Equation (2.7), the *Precision* as in Equation (2.8), the F_1 -score or Dice coefficient, as in Equation (2.9), and the *Specificity*, as in Equation (2.10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.6)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.7)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.8)$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (2.9)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2.10)$$

2.1.7.2 Semantic Segmentation

Semantic segmentation can be seen as a pixelwise or voxelwise classification problem, but more specialized metrics are often required. Main quality measures in this context have been widely adopted in challenges as SLIVER07 and LiTS [126, 127], and they represent the gold

Table 2.1 *Binary classification confusion matrix*. N and P stand for Negative and Positive, respectively. TP , TN , FP and FN are the number of True Positives, True Negatives, False Positives and False Negatives, respectively.

		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

standard in segmentation tasks for Medical Imaging. Two fundamental categories of metrics exists: those based on the volumetric overlap, inherently linked to the pixelwise or voxelwise confusion matrix, and those based on surface distances.

The most adopted quality measures founded on volumetric overlap include the Jacard Index $J(\cdot, \cdot)$, also called Intersection over Union (IoU), the Volumetric Overlap Error $VOE(\cdot, \cdot)$, and the Sørensen–Dice Coefficient $DSC(\cdot, \cdot)$, defined, respectively, in Equation (2.11), Equation (2.12) and Equation (2.13). In all the definitions involved in the quality measures, B denotes the binarized predicted segmented volume (obtained by thresholding predicted probability maps) and G the ground truth volume; the cardinality operator for a set is denoted as $|\cdot|$.

$$J(B, G) = \frac{|B \cap G|}{|B \cup G|} \quad (2.11)$$

$$VOE(B, G) = 1 - J(B, G) \quad (2.12)$$

$$DSC(B, G) = \frac{2 \cdot |B \cap G|}{|B| + |G|} \quad (2.13)$$

The concept of overlap behind $DSC(B, G)$ and $J(B, G)$ can be generalized, resulting in the Tversky Index $T_{\alpha, \beta}(B, G)$, defined as:

$$T_{\alpha, \beta}(B, G) = \frac{|B \cap G|}{|B \cap G| + \alpha |B - G| + \beta |G - B|} \quad (2.14)$$

where it is worth noting that $T_{0.5, 0.5}(B, G)$ corresponds to $DSC(B, G)$, and $T_{1, 1}(B, G)$ is equivalent to $J(B, G)$.

A more approximate indication about the relative difference between the volumes is the Relative Volume Difference (RVD), which can be defined as in Equation (2.15).

$$RVD(B, G) = \frac{|B| - |G|}{|G|} \quad (2.15)$$

For what concerns quality measures designed for assessing external surface distances, it is worth mentioning the Maximum Symmetric Surface Distance (*MSSD*), the Average Symmetric Surface Distance (*ASSD*), and the Root Mean Square Symmetric Surface Distance (*RMSD*). These measures are particularly useful for applications such as surgical planning and image-guided surgery, where it is vital to obtain a suitable prediction of the mesh of the considered organ.

To properly define these distances, a metric space (X, d) has to be defined. In this metric space, X is a 3D Euclidean space and d is the Euclidean distance over the space. Then, let $S(B)$, $S(G)$ in X be, respectively, the external surfaces of the B and the G volumes. A distance function between any two non-empty sets $S(B)$ and $S(G)$ of X , also known as one-sided Hausdorff distance, $h(S(B), S(G))$, can be defined as in Equation (2.16).

$$h(S(B), S(G)) = \sup_{s_B \in S(B)} \left\{ \inf_{s_G \in S(G)} d(s_B, s_G) \right\} \quad (2.16)$$

Then, the *MSSD*, also known as bidirectional Hausdorff distance, can be defined as in Equation (2.17), whereas the *ASSD* can be defined as in Equation (2.18) and *RMSD* as in Equation (2.19).

$$MSSD(B, G) = \max \{h(S(B), S(G)), h(S(G), S(B))\} \quad (2.17)$$

$$ASSD(B, G) = \frac{1}{|S(B) + S(G)|} \left(\sum_{s_B \in S(B)} d(s_B, S(G)) + \sum_{s_G \in S(G)} d(s_G, S(B)) \right) \quad (2.18)$$

$$RMSD(B, G) = \sqrt{\frac{1}{|S(B) + S(G)|}} \cdot \sqrt{\sum_{s_B \in S(B)} d(s_B, S(G))^2 + \sum_{s_G \in S(G)} d(s_G, S(B))^2} \quad (2.19)$$

A metric based on *MSSD*, which is also adopted in challenges (<https://structseg2019.grand-challenge.org/Evaluation/>) is *95%MSSD*, referring to the 95th percentile of *MSSD*, with the purpose to eradicate the impact of a small subset of outliers.

2.1.7.3 Object Detection

Reference metrics used for evaluating object detection models are based on object detection challenges as PASCAL VOC (<http://host.robots.ox.ac.uk/pascal/VOC/>), Google Open Images (<https://opensource.google/projects/open-images-dataset>), and COCO (<https://cocodataset.org/>). In general, the performance metrics used in these challenges offer a global level evaluation, estimating the performances of the model in the whole dataset. The adoption of global metrics makes benchmarking much simpler, but it does not provide insights on how and why the mistakes have been made.

In order to define object detection metrics, it is necessary to properly define a detection first. Given two bounding boxes A and B , it is possible to define Intersection over Union (IoU) and Intersection over Minimum (IoM). IoU is the ratio between the intersection and the union of the bounding boxes areas, as defined in Equation (2.20).

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2.20)$$

As for semantic segmentation metric definitions, $|\cdot|$ denotes the set cardinality operator. IoU values lie in the range $[0, 1]$, where 1 indicates a perfect match. A predicted object matches with a ground truth object when IoU between them is above a certain threshold (a common choice for the threshold is 0.5).

On the other hand, the IoM between two bounding boxes A and B is the ratio between the intersection and the minimum of their areas, as reported in Equation (2.21).

$$IoM(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (2.21)$$

IoM values lie in the range $[0, 1]$, where 1 indicates a perfect match.

Widespread evaluation metrics for object detection are Average Precision (AP), which can be mainly defined as the area under the precision–recall curve, and mean Average Precision (mAP), that is AP averaged over all classes. A naïve implementation of AP is described in Equation (2.22):

$$AP = \int_0^1 p(r) dr \quad (2.22)$$

Anyway, it is important to note that AP is usually calculated (e.g., PASCAL VOC) by adopting the average interpolated precision value of the positive examples [128]. The dependence of precision and recall from confidence c can be explicated using the notation $p = P(c)$

and $r = R(c)$. Recall $R(c)$ is the fraction of objects detected with confidence of at least c . Precision $P(c)$ is the fraction of detections that are correct:

$$P(c) = \frac{R(c) \cdot N_j}{R(c) \cdot N_j + F(c)} \quad (2.23)$$

In Equation (2.23), N_j is the number of objects in class j and $F(c)$ is the number of incorrect detections with at least confidence c .

Mean Average Precision (mAP) for K classes can be calculated as reported in Equation (2.24):

$$mAP = \frac{1}{K} \sum_{j=1}^K AP_j \quad (2.24)$$

2.1.7.4 Image Similarity

Other quality measures that have to be noted are Peak-Signal-to-Noise Ratio ($PSNR$) and Structural Similarity Index Measure ($SSIM$), which are especially useful in assessing image degradation (as blurring, noise, or JPEG compression). Horé et al. derived a mathematical relationship between them [129]. Given a reference image f and a test image g , both having size $(M \times N)$, and given $MSE(f, g)$ defined as in Equation (2.25):

$$MSE(f, g) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (f_{ij} - g_{ij})^2 \quad (2.25)$$

it is then possible to introduce $PSNR(f, g)$ as in Equation (2.26):

$$PSNR(f, g) = 10 \log_{10} \left(\frac{p^2}{MSE(f, g)} \right) \quad (2.26)$$

where P is the peak value of the signal.

Instead, the definition of $SSIM$, reported in Equation (2.27), is more complex, since it involves the combination of three factors: loss of correlation $s(f, g)$, luminance distortion $l(f, g)$ and contrast distortion $c(f, g)$. For more details about $SSIM$ refer to [129].

$$SSIM(f, g) = l(f, g) c(f, g) s(f, g) \quad (2.27)$$

2.1.8 Explainable AI Methods

Two families of methodologies have a prominent role in interpretability and explainability: perceptive interpretability and mathematical interpretability [130]. Perceptive XAI involves the straightforward visualization of the most important features affecting the final model outputs, allowing to study feature level classification patterns of the DL network topologies. On the other side, mathematical interpretability furnishes insights concerning the developed models, portraying the features adopted to realize the predictions, often realizing cluster maps.

2.1.8.1 Perceptive XAI

Two of the most widespread methods for carrying out perceptive XAI explanations are Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME) [130]. They are post-hoc techniques, which work on existing trained models without modifying them, that can be exploited to interpret the decisions performed by DL models. The remainder of this section details the behavior of the Grad-CAM and LIME techniques.

Grad-CAM Grad-CAM can be considered as a method which rely on back-propagation, since the algorithm performs both forward-pass and backward-pass through the network topology, exploiting partial derivatives of the activations [131]. The original CAM method needed a particular design of the DL model under consideration. Unlike its predecessor, Grad-CAM offers a generalization which can be employed without altering the underlying CNN architecture [132].

The Grad-CAM method generates a class activation heatmap for a given network layer and input image. This means that, for a given class, the Grad-CAM tries to realize a comprehensible representation of the decision making mechanism of the underlying DL model as an heatmap that shows the feature importance. In the last layers of CNN architectures, neurons examine semantic information related to specific classes. Grad-CAM is capable to exploit the gradient flow of one of these layers to decipher feature importance information that can be depicted on a saliency map. At the end, usually, the obtained representation is portrayed with a jet color scheme. Red color depicts saliency zones, whereas blue the background.

LIME Another well-known technique for explainability is LIME. It is a model-agnostic method, which can be used for any DL architecture. Its name derives from its properties:

- **Local** – the behavior of the network is explained by estimating its local behavior;
- **Interpretable** – the output of the LIME method is useful to get an insight on the model functioning for a human user;
- **Model-Agnostic** – the LIME technique does not depend on the model under consideration, but can be applied for every existing black-box model.

When used in classification setups, LIME can be explained as follows. A patch of pixels which similar characteristics (superpixels) is taken from the original input image, after having generated a linear model, and used to obtain samples. Computation of superpixels in the image is carried out with the quick-shift algorithm. Then, perturbation images are obtained and predictions are made on those.

Lastly, a saliency map is superimposed to the image, so to highlight important regions, i.e. superpixels that contributed the most in the classification. A common convention is to denote the positively contributing features with green color; the negatively contribution features are portrayed in red color. The number of top contributing features to show is a parameter that can be adjusted.

2.1.8.2 Mathematically Explained XAI

Two widely employed and useful techniques, t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), for achieving mathematical interpretability are described in this section. Those techniques allow to perform dimensionality reduction of the features extracted by CNN and other DL topologies, permitting to visualize clusters in a low-dimensional space.

The main purpose of t-SNE and UMAP concerns data visualization. However, it has to be noted that there is a difference in the interpretation of the inter-cluster distance. Indeed, UMAP can retain both local and global structure in the data, while t-SNE focuses on maintaining the local structure. Therefore, UMAP is a more appropriate technique for assessing the dissimilarity and inter-cluster distance.

t-SNE The t-SNE methodology is a modification of the SNE technique, which enables the visualization of data from high-dimensionality spaces into low-dimensionality spaces (for practical reasons, of 2 or 3 dimensions) [133]. Compared to the SNE method, it improves on two aspects:

1. the cost function optimization – a Student-t distribution is employed to change the SNE cost function (symmetrized), so to compute data similarity in the low-dimensional space.
2. the "crowding problem" – a heavy-tailed distribution is exploited in the low-dimensional space.

UMAP The UMAP methodology offers a nonlinear mapping for dimensionality reduction, grounded on three assumptions [134]:

1. the data has a uniform distribution on an existing manifold;
2. the manifold topological structure should be preserved;
3. the manifold is locally connected.

The UMAP method consists of two main stages. The first involves learning a manifold structure in a high-dimensional space, whereas the second concerns finding a related representation in the low-dimensional space. The initial step of the first stage consists in finding the nearest neighbors for the datapoints, by taking advantage of the Nearest-Neighbor-Descent algorithm.

Thus, a graph is built by linking the previously identified neighbors. Since data are uniformly distributed among the manifold, the space between datapoints fluctuates depending on regions where data are sparse or denser. The concept of *edge weights* can be introduced exploiting this assumption. It consists in determining, for every point, the distance to its nearest neighbor. Though, a problem of disagreeing edges can happen.

2.2 Radiomics and Deep Learning in Radiology

In the clinical field, the process of volumetric organs quantification plays a pivotal role. This process is required for diagnosing pathologies and monitoring their evolution over time, besides surgical planning setups. Volumetric imaging acquisition techniques, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), allow for three-dimensional delineation of organs in a non-invasive way. In clinical practice, the gold standard often relies on manual annotation performed by radiologists, which is time-consuming and error-prone, having to label each two-dimensional slice of the volumetric acquisition

manually. Moreover, there are concerns about inter- and intra-operator variability of the resulting segmentation [135].

The research community made a considerable effort to design and develop reliable segmentation methods, explicitly focusing on anatomical sites having a morphological structure that could vary over time, both for physiological and pathological reasons, such as kidneys or liver. Thanks to recent advances in Artificial Intelligence (AI) algorithms, DL methodologies and architectures, such as CNNs, have become very popular in different application domains, including the medical context. The peculiarity of DL techniques is that these approaches do not require to design procedures for extracting features, which is well-known to be a challenging and time-consuming process, also requiring a very profound knowledge of the application domain [136, 137]. In fact, before the CNN outbreak, different approaches, including the Chan-Vese model [138], neural network-based approaches [139, 140], and discriminant-based approaches [141, 142], were adopted for features-based image analysis.

The drawback is that, in order to train supervised DL models, there is a need for a huge quantity of training data. This problem can be partially solved by techniques as Transfer Learning or Self-Supervised Learning, but in the medical imaging domain there is a huge effort in realizing datasets and challenges which are publicly available and count on international collaborations for providing the data, thus allowing to ease the initial annotation process. After that models have been trained, human-in-the-loop methodologies can be further exploited to delineate even more data [143, 144].

Focusing on research applications dealing with CNNs applied on Medical Imaging, it is worth noting that more than 8000 documents on Scopus and more than 80000 results on Google Scholar have emerged in the last five years, with a continuously growing trend, as reported in Figure 2.3. For what concerns the number of publications in Medical Imaging, Scopus has been queried with “*TITLE-ABS-KEY ((convolutional AND neural AND network) OR (deep AND learning) AND (medical AND imaging))*”, whereas Google Scholar has been searched with “*convolutional neural network medical imaging*”. For what regards the publications in Deep Learning for segmentation of liver, kidney and spleen, Scopus has been queried with “*TITLE-ABS-KEY ((“semantic segmentation” OR “deep learning”) AND (“\$organ”) AND (“CT” OR “MRI”))*”, where *\$organ* refers to liver, kidney or spleen.

The adoption of Deep Learning strategies allowed the development of many automatic CDSSs from medical images [18, 19, 36, 145–156]. CDSSs have an essential role also for what concerns the automatic delineation of organs from volumetric imaging acquisitions [49, 157–169]. In particular, the abdominal area has been the subject of several studies

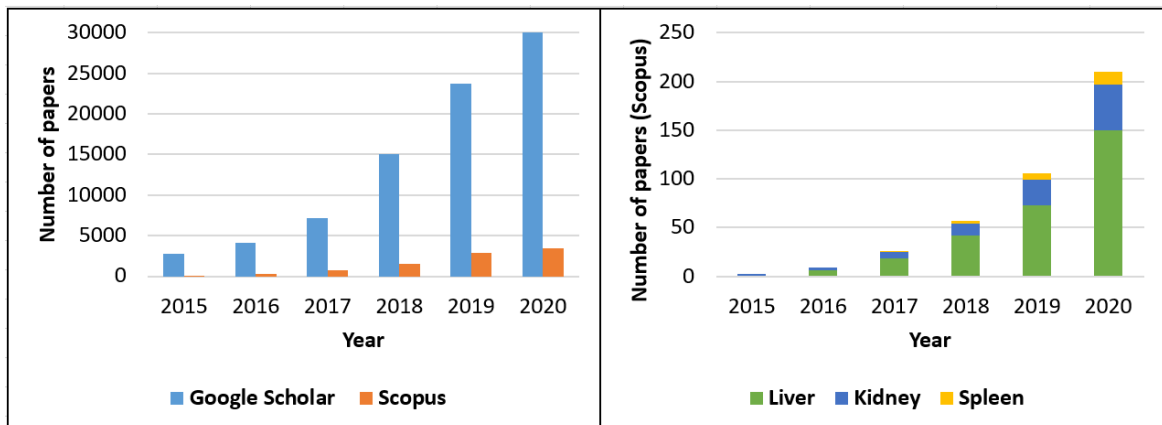


Fig. 2.3 *Deep Learning Publications in Medical Imaging*. Left: number of publications of Deep Learning in Medical Imaging on Scopus and Google Scholar. Right: number of publications per year of Deep Learning for segmentation of liver, kidney and spleen.

for applications like radiotherapy treatment planning, volume measurement and computer-assisted surgery [170].

Traditionally, the visual assessment of medical images for detection, characterization, and monitoring of pathological conditions is performed by trained radiologists. Artificial Intelligence methodologies can improve the traditional workflow by providing quantitative measurements for radiographic characteristics by exploiting pattern recognition capabilities in imaging data in automatic or semi-automatic ways. These characteristics can result in significant medical field progress, especially when considering oncological applications [171].

Radiology is intrinsically a data-driven specialty; thus, data processing methodologies (as those based on Deep Learning) are particularly relevant in this clinical domain [172]. Indeed, several works have already proven that DL is a suitable approach for solving different tasks, including disease or lesion detection, classification, quantification, and segmentation [17, 56, 145, 157, 158, 171–185].

One fundamental question that arises from adopting DL for medical applications concerns the difference between fine-tuning and full training. Performing the training process from scratch is difficult, given the vast amount of labeled training data required. Labeled natural images collections, like ImageNet [186–191], can pose the basis for implementing fine-tuning mechanisms. However, the central question is whether labeled natural images can provide a reliable source of information for discriminating against medical images. Semi-supervised image annotation can also be exploited [192]. According to Tajbakhsh et al., by employing a

pre-trained CNN with fine-tuning, it is possible to outperform CNN trained from scratch; moreover, fine-tuned CNNs are more robust to the size of training sets [175].

Even if all the points so far look very promising, DL techniques suffer from some limitations. Besides the need for a lot of consistent data for training (worsen by some companies not sharing their intellectual properties, keeping proprietary their datasets [193]), there are also other points to note. First, ANNs can be “statistically impressive, but individually unreliable”, making mistakes that humans would not [194]. Moreover, ANNs, also in the Medical Imaging domain, can be subject to adversarial attacks, i.e. researchers could modify images in such a not detectable to the human eye but able to trick DL image classifiers [195–198].

DL-based methodologies in abdominal radiology have a fundamental role for a wide array of applications, including diagnosis, prognosis, diseases monitoring, radiotherapy, and surgical planning. Furthermore, in all these contexts, the development of CAD solutions based on DL pipelines for organs segmentation can ease the burden of the radiologist. The segmentation of these abdominal organs is developed in several studies as reported in [199–202]. The main challenges in the segmentation of liver, kidney, and spleen include variations in morphology, location, orientation, and intensity values, both among more patients than in the same person considered in different time spans [203]. Moreover, the use of different scanners may result in variable quality of obtained images.

2.2.1 Datasets and Challenges

Biomedical image segmentation is becoming more and more important in the biomedical research field, resulting in widespread efforts for producing publicly available datasets. Increasingly, organizers of medical imaging challenges publicly release this data. Some of the most important dataset collections are:

- **Medical Segmentation Decathlon** [204]. Semantic segmentation approaches are incredibly functional for the automatic segmentation of Regions of Interest (ROIs), especially in the Medical Imaging community. The role of this collection of datasets is to provide the users with annotated medical image data of different clinically relevant sites. Data is released under an open-source license. It also sets a comprehensive benchmark employable to assess the relevance of proposed methods. Among the ten datasets provided, Task03 (Liver), Task08 (Hepatic Vessels), and Task09 (Spleen) are detailed in the remainder of this section.

- **MedPix** [205]. It is a free open-access online database with about 59,000 images belonging to more than 12,000 patients. The main target audience comprises physicians, nurses, health professionals, and students interested in medical knowledge. Contents are classified by disease location (organ system); type of pathology; profile of each patient; image classification and captions. This results in a data collection in which any search can be made by diagnosis, organ, patient symptoms and signs, image modality and description, contributing authors and keywords, among other options.
- **Grand Challenge** [206]. It is a platform for easing the end-to-end development of machine and deep learning algorithms in the biomedical imaging domain. In June 2022, it hosts more than 300 challenges, with roughly 69,000 subscribed users, but it is growing even more rapidly.
- **MICCAI Challenges** [207]. The MICCAI Society annually holds the MICCAI International Conference, a leading conference in the Medical Imaging domain. During this conference, a few relevant challenges are organized. Starting from MICCAI 2021, the process of challenges organization has become even more rigorous, with the development of a structured challenge submission system (<https://www.biomedical-challenges.org/miccai2021/>), allowing to overcome many of the issues related to these challenges, as reproducibility and interpretation of the results and ranking of the proposed algorithms [208].
- **The Cancer Imaging Archive (TCIA)** [209]. Born from the National Cancer Institute and Washington University (Saint Louis) effort, it is a very large database of cancer images collections. In the first year of activity, it accumulated over 20 collections and 3 million images. TCIA is an open-source and open-access information resource that sustains didactic, research, and development activities based upon advanced Medical Imaging of cancer.

In the following of this section, the considered datasets and challenges are split into two categories:

- **Single-Organ Segmentation**, which refers to publicly available datasets that contain labels for only one organ among liver, kidney, and spleen.
- **Multi-Organ Segmentation**, which refers to publicly available datasets that contain labels for different organs, among which at least one is liver, kidney, or spleen.

2.2.1.1 Single-Organ Segmentation

The two most important benchmarks for **Liver Segmentation** are SLIVER07 [126] and Liver Tumor Segmentation Benchmark (LiTS) [127]. Further, there is also the Medical Segmentation Decathlon Task 08 (MSD 08). Note that MSD Task 03, from the same collection, hosts the same data of LiTS.

In SLIVER07, there is a comparison between 10 automatic and 6 interactive methods for Liver Segmentation from contrast-enhanced CT images, based on the workshop “MICCAI 2007 Grand Challenge”. The considered algorithms include atlas registration, level-sets, graph-cuts, rule-based systems, and statistical shape models. SLIVER07 dataset contains 20 CT scans for training, for which liver masks are publicly available, and 10 test CT scans. According to the specific machine and protocol adopted, CT scans of SLIVER07 have an x/y pixel spacing which belongs to the range 0.55-0.8 mm, and a thickness ranging from 1 mm to 3 mm.

The LiTS challenge has been organized in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 and the MICCAI 2017 International Conference [127]. Twenty-four approaches for liver and liver tumor segmentation have been considered. According to the LiTS organizers, the best-ranked algorithms are based on U-Net [72] architectures (a widely adopted CNN in the Medical Imaging domain, detailed in Section 2.1) and have common properties, such as exploiting cascaded approaches, residual connections, and 2.5D architectures. The LiTS dataset contains 201 CT abdominal acquisitions. The training set is composed of 131 CT scans, whereas the test set contains 70 CT scans. Liver and liver tumor masks are publicly available for the training set only. Pixel spacing, in x/y-direction, belongs to the range 0.56-1.0 mm, whereas slice distance ranges between 0.45 mm and 6.0 mm, with the number of slices varying from 42 to 1026. Note that LiTS training dataset contains the public 3D-IRCADb dataset.

The Medical Segmentation Decathlon Task 08 dataset consists of 443 portal venous phase CT scans [204]. In this dataset, the x/y pixel spacing belongs to the range 0.57mm to 0.9mm, whereas the thickness goes from 2.5mm to 5.0 mm. For this dataset, there are also publicly available CT scans annotations for liver parenchyma and segments, made by expert radiologists for a study carried out by Tian et al. from Lenovo Research [210].

The main publicly available datasets for **Kidney Segmentation** include the 2019 Kidney Tumor Segmentation Challenge (KiTS19) and the 2021 Kidney and Kidney Tumor Segmentation Challenge (KiTS21) [211–213].

The KiTS19 challenge dataset is a collection of 300 CT scans, with related segmentation masks and treatment outcomes, for the same number of patients treated with partial or radical

nephrectomy in the years 2010-2018. The train set is composed of 210 CT scans, for which the label data are publicly available. The remaining 90 CT scans are used for objectively assessing the proposed methods [212]. The competition has been held in conjunction with the 2019 MICCAI International Conference. More than 100 teams joined the KiTS19 challenge, developing automated systems for predicting kidney and kidney tumor segmentation for the test set [213].

The KiTS21 challenge (<https://kits21.kits-challenge.org/>) has been accepted for the 2021 MICCAI International Conference, and it will incorporate more data compared to KiTS19, allowing to overcome the limitations of the previous challenge by integrating acquisitions from different geographical locations, also at different times. Even if the challenge organizers wanted to include labels for ureters and renal vessels, they finally decided to not include them anymore in the challenge. With respect to the KiTS19 challenge, it is worth noting that annotations for cysts, defined as kidney masses radiologically determined to be cysts, have been added [211].

Regarding the publicly available single-organ datasets for **Spleen Segmentation**, only the Medical Segmentation Decathlon Task 09 (MSD 09) dataset has been found. MSD 09 dataset consists of 61 portal venous phase CT scans (41 constituting the training set, and the remaining 20 as test set) from patients who underwent chemotherapy treatment for liver metastases at the Memorial Sloan Kettering Cancer Center (New York, NY, USA) [204].

Table 2.2 reports a summary of the datasets for the single-organ segmentation.

Dataset	Annotation	x/y spacing	Thickness	Number of slices	Number of labeled images	Publication Year
SLIVER07	L	0.55-0.8mm	0.45-6mm	64-394	20 (+10)	2009
LiTS, MSD 03	L, LT	0.56-1mm	1-3mm	74-987	131 (+70)	2019
MSD 08	LV, LT	0.57-0.9mm	2.5-5mm	24-251	443 (+140)	2019
	L, LS				193	
KiTS19	K, KT	0.44-1.04mm	0.5-5mm	29-159	210 (+90)	2019
KiTS21	K, KT, KC	N/A	N/A	N/A	N/A	2021
MSD09	S	0.61-0.98mm	1.5-8mm	31-168	41 (+20)	2019

Table 2.2 *Single-Organ Datasets Summary*. The number of labeled images is reported as “training labels (+ testing labels)”. Annotation acronyms: L stands for Liver; LT stands for Liver Tumors; LV stands for Liver Vessels; LS stands for Liver Segments; K stands for Kidney; KT stands for Kidney Tumors; KC stands for Kidney Cysts; S stands for Spleen.

2.2.1.2 Multi-Organ Segmentation

Among the most important collections for Multi-Organ Segmentation which include at least one between the liver, kidney, and spleen, there are the 3D-IRCADb dataset (<https://www.ircadb.com/>):

[//www.ircad.fr/research/3d-ircadb-01/](http://www.ircad.fr/research/3d-ircadb-01/)), the CT-ORG dataset [214, 215], the CHAOS dataset [216–218] and a dataset from Gibson et al. [159].

The 3D-IRCADb dataset (<https://www.ircad.fr/research/3d-ircadb-01/>) is composed of 20 CT scans. The liver, but also vessels and other abdominal organs (among which there are kidneys and spleen) are annotated. The 3D-IRCADb dataset contains CT scans whose axial-plane resolution varies from 0.56 mm to 0.81 mm, whereas resolution along the z-axis spans from 1 mm to 4 mm. The number of slices ranges between 74 and 260. The 20 CT scans of 3D-IRCADb come from 10 women and 10 men; the number of patients with hepatic tumors is 75% of the overall dataset.

CT-ORG dataset [214, 215], from TCIA [209], comprises 140 CT scans with five organs labeled in 3D: lung, bones, liver, kidneys and bladder. In the minority of scans with the brain, annotations are provided for it. Most images contain liver lesions, both malignant and benign. There are also scans that exhibit metastatic disease in other organs, such as bones and lungs. The vast majority of images, 131 out of 140 CT scans, comes from LiTS [127].

CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation (<https://chaos.grand-challenge.org/>) dataset consists of 20 training and 20 testing cases in the CT dataset. MRI dataset contains 20 training and 20 testing cases with T1-Dual and T2 SPIR sequences. Train data contains both DICOM images and their ground truth masks. The testing set only contains DICOM images. In CT cases, only livers were annotated. Instead, in MRI cases, livers, left/right kidneys, and spleens were annotated [216–218].

Gibson et al. shared a dataset composed of 90 abdominal CT scans which delineate multiple organs: spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum [159]. The abdominal CT images and some of the reference segmentation are derived from two datasets: TCIA Pancreas-CT dataset [209, 219, 220] and Beyond the Cranial Vault (BTCV) Abdomen data set [221, 222]. The dataset has been published on a public repository (<https://zenodo.org/record/1169361#.X6U6lWhKiUk>).

Table 2.3 reports a summary of the datasets for the multi-organ segmentation, whereas Figure 2.4 shows an image representing examples for liver, kidney and spleen from some of the different datasets listed in Table 2.2 and Table 2.3.

2.2.2 Segmentation Workflow

The general workflow implemented for segmentation in radiology may be delineated as follows:

Dataset	Annotation	x/y spacing	Thickness	Number of slices	Number of labeled images	Publication Year
3D-IRCADb	L, LT, V, K, S, etc.	0.56-0.81mm	1-4mm	74-260	20	2010
CHAOS [216–218]	L, K, S	0.54-2.03mm	2-8mm	26-294	40 (+40)	2019
CT-ORG [214, 215]	L, K, Lu, Bo, Bl, Br	0.56-1.37mm	0.7-5mm	74-987	140	2019
Dataset [159]	S, Kl, G, E, L, St, P, D	0.59-0.98mm	0.5-5mm	85-466	90	2018

Table 2.3 *Multi-Organ Datasets Summary*. The number of labeled images is reported as “training labels (+ testing labels)”. Annotation acronyms: L stands for Liver; LT stands for Liver Tumors; V stands for Vessels; Ks stands for Kidneys (Kl stands for left Kidney); S stands for Spleen; Bo stands for Bone; Bl stands for Bladder; Br stands for Brain; Lu stands for Lung; G stands for Gallbladder; E stands for Esophagus; St stands for Stomach; P stands for Pancreas; D stands for Duodenum.

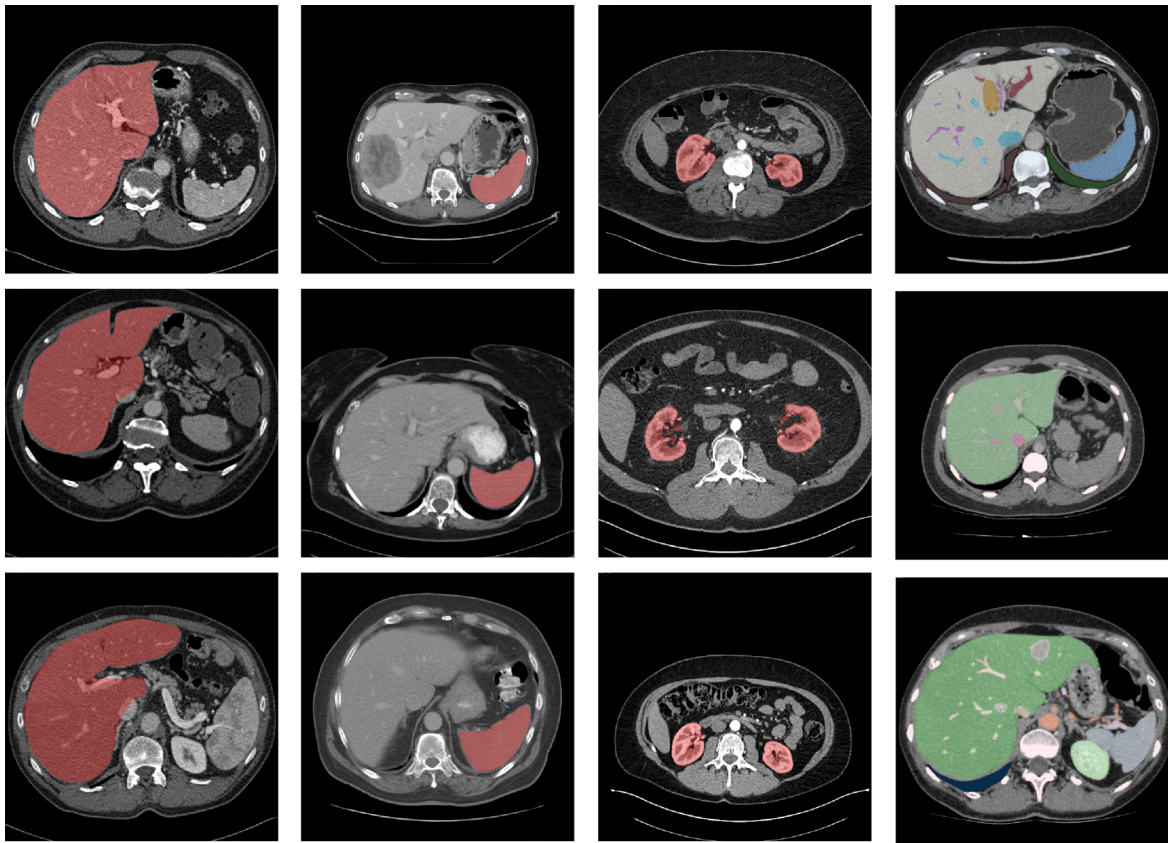


Fig. 2.4 *Liver, Kidney, and Spleen image examples*. Data from LiTS [127], KiTS19 [213], MSD 09 [204] and 3D-IRCADb. Twelve slices from different patients are reported. The first column contains liver examples, the second column spleen samples, the third column kidney examples and the last column multi-label examples.

1. *Data Preparation*. This step could be very dependent on the specific research which is carried out. If a publicly available dataset is exploited, typically, the only operation required is to obtain permission from dataset creators and download it. If data has to be collected from a local hospital, there are other aspects to consider, such as obtaining ethical approval and producing high-quality ground truth segmentation masks.
2. *Data Augmentation* and/or *Pre-processing*. These two steps have been grouped since they could involve the same operations. Moreover, it is possible to perform augmentation and pre-processing in different orders.
 - (a) *Data Augmentation*. There exist numerous data augmentation libraries available for natural and medical images [223–225]. Typically, augmentations can be categorized into:
 - i. Pixel-level transformations, as those intensity-based, such as rescaling, z-score normalization, histogram standardization, random noise, random blurring.
 - ii. Spatial-level transformations, such as geometric, morphological, and textural ones, e.g. crop, flip, grid distortion, affine transformations, perspective transformations, elastic transformations, resize, rotation. Moreover, recently GANs have started to show promising results in augmenting biomedical image datasets [29, 96, 97, 226].
 - (b) *Pre-processing*. Classical pre-processing steps may include clipping and rescaling, z-score normalization, histogram matching, and histogram equalization. In other cases, where there are artifacts that could impact the quality of the learning model, more advanced steps have to be implemented to obtain coherent input data.
3. *Learning Model*. Since the main focus of the thesis work is on learning models (specifically on Deep Learning ones) for CAD and CDSS, other classical techniques that could have been used (e.g., region growing, level set) are not considered. Among DL-models, the most frequently exploited ones for image segmentation are:
 - (a) FCN or E/D (encoder-decoder) architectures, which have been detailed in Section 2.1.2.
 - (b) SNN (Siamese Neural Network), which consists of two identical ANNs that learn hidden representations of an input vector for their comparison using a distance

measure [227]. This idea stems from a system for signature verification developed some decades ago [228], but they are recently gaining attention for applications such as tracking [229], one-shot learning [230], and medical image segmentation [231].

- (c) GANs, which have been already introduced in Section 2.1.4.
- 4. *Post-processing*. It is an optional additive step that could refine the segmentation results. Techniques like graph-cut and CRF (introduced in Section 2.1.6) have proven their usefulness in the analyzed papers which adopted them. Sometimes, also more basilar methodologies, like connected components labeling and morphological operators, may lead to improvements [56].
- 5. *Evaluation*. Typically, during the training loop, intermediate results are evaluated on both the training set, for assessing the training loss and performing backpropagation, and on a subset kept out for validation, which can be used to implement early stopping or for selecting the best model. After the training process, the obtained model is assessed on both internal and external validation sets.

A conceptual representation of the presented workflow is delineated in Figure 2.5.

2.2.3 Liver Parenchyma Segmentation

The importance of liver (parenchyma) segmentation in the medical domain has a long tradition, with the first important “Grand Challenge” held in MICCAI 2007 [126]. Liver segmentation is the first fundamental step for computer-based surgery planning of interventions as resections of tumors, living donor transplantations, or minimally invasive surgery [232, 233]. Furthermore, it is also useful for monitoring and diagnosing liver diseases [126, 127, 234, 235].

CT scan images are typically acquired for these reasons. Difficulties incurring in the automatic delineation of liver parenchyma include the presence of adjacent organs with similar intensities, such as heart and stomach, or subcostal fat of the rib cage. Moreover, pathologies, such as cirrhosis, partial liver resections, or wide tumors, pose further issues [126].

Subsequently to liver segmentation, liver lesion classification can be efficiently approached by CNNs [236]; furthermore, it can be aided by adding synthetic data exploiting GANs, resulting in an increase in recall and specificity [226]. Besides parenchyma segmentation, another important step concerns vessels delineation. In modern hepatic surgery, the

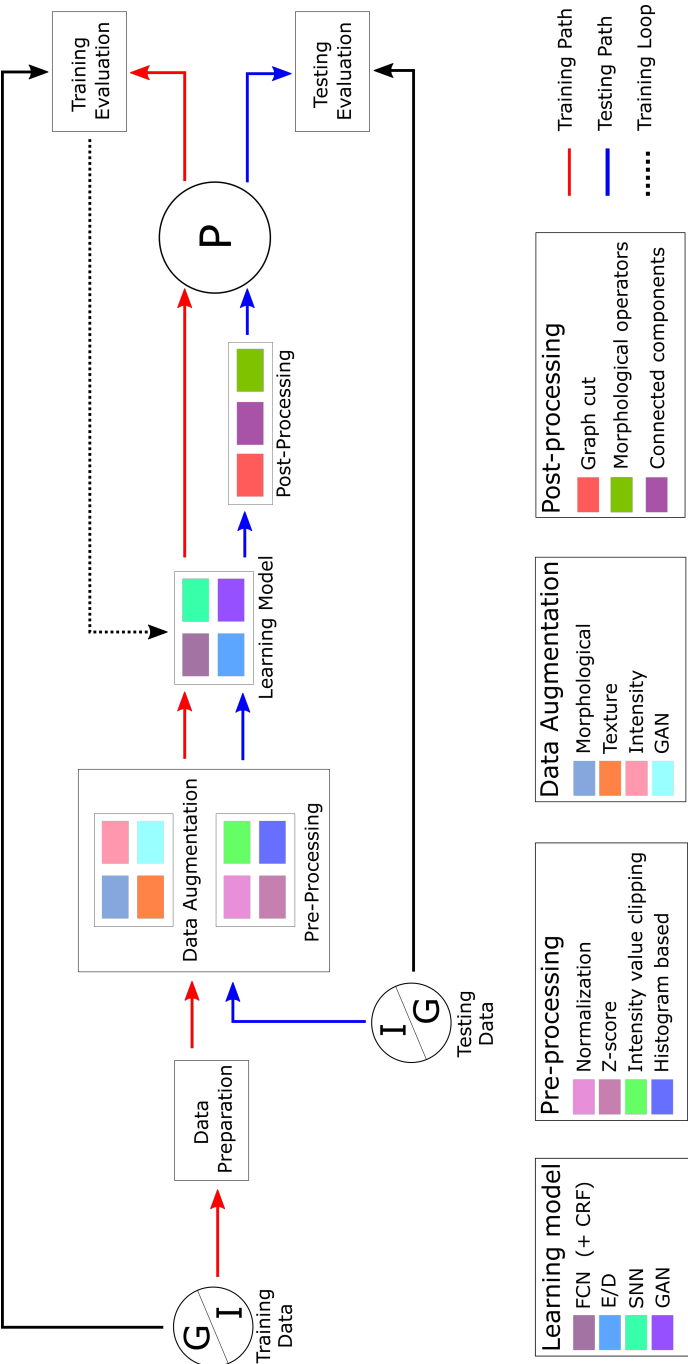


Fig. 2.5 Segmentation Workflow: A Schematic Representation. The main steps involved are Data Preparation, Data Augmentation and Pre-processing, Learning Model, Post-processing, and Evaluation. I denotes the original image (e.g. CT scan or MRI), G the ground truth segmentation mask, and P the model prediction.

correct identification of hepatic vessels is fundamental since they represent the anatomic borders of hepatic segments (according to the model of Couinaud [237]). Segmentectomies, based on the correct identification of the vascular landmarks, can avoid the unnecessary removal of healthy liver parenchyma, also reducing the complications of most extensive resections [238]. In literature, there are different classical approaches aimed at liver vessels segmentation. Region-based approaches, *k*-means clustering, morphological operators, context-based voting algorithms, oriented flux symmetry, and graph-cut have been exploited for the task [239–245]. More recently, DL methodologies are emerging for liver vessels segmentation too [56, 246].

Existing DL approaches for liver segmentation can be categorized into three main categories: pure CNNs [85, 87, 88, 247, 248], CNNs combined with CRFs [147, 249, 250] or other post-processing techniques [56, 251–253], and adversarial loss set-ups [254, 255]. Approaches that include SSL have been reported in Section 2.1.5.

2.2.3.1 Pure CNN approaches

Zhou et al. presented UNet++, a deeply supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. The authors tested UNet++ on a variety of tasks, including liver segmentation in abdominal CT scans, showing improvements in terms of IoU with respect to the standard U-Net architecture [85, 88].

Li et al. proposed a novel architecture, the H-DenseUNet (Hybrid Densely Connected U-Net), for liver and liver tumor segmentation from CT scans. The authors realized a hybrid feature setup, where a 2D network is exploited for intra-slice feature extraction, and a 3D network is adopted for the hierarchical reconstruction of volumetric context. The method has been evaluated on the LiTS challenge [127] and 3D-IRCADb dataset [87].

Ben-Cohen et al. compared different variants of FCN, based on the VGG-16 [41] model, for liver and liver tumor segmentation from CT scans. They conclude that the obtained results may be clinically relevant [247, 248].

2.2.3.2 CNNs combined with post-processing

Christ et al. [147] proposed a method for automatic liver and liver tumor segmentation by exploiting Cascaded Fully Convolutional Neural Networks (CFCNs), training two FCNs, and dense 3D CRFs. The models were trained on the abdominal CT scans of the 3D-IRCADb dataset. The authors concluded that 3D CRFs are beneficial as post-processing, resulting

in an improvement of the quality measures employed (related to the original SLIVER07 challenge [126]).

Dou et al. introduced 3D DSN, a 3D FCN trained with a deep supervision mechanism, which exploits intermediate feature maps, after upsampling, to obtain an extra-dense prediction. In the loss function, the components of these extra prediction maps have a decaying weight. CRF is applied as post-processing for contour refinement, increasing the segmentation accuracy. The authors tried different tasks, among which there is liver segmentation from 3D CT scans [249].

Rafiei et al. developed a 3D to 2D Fully Convolutional Network (3D-2D-FCN) to speed up the recognition of trauma areas since liver segmentation in emergencies is a vital task. The authors showed a performance gain with the CRF adoption compared to the CNN-only model [250].

Lu et al. presented a DL algorithm with graph-cut refinement to perform automatic liver segmentation from CT scans. A 3D CNN is exploited to perform liver detection and probabilistic segmentation; then, the initial segmentation is refined via graph-cut, taking advantage of the previously learned probability maps. The approach has been validated on volumes taken from SLIVER07 [126] and 3D-IRCADb datasets. The authors concluded that the obtained results are efficient and accurate for hepatic volume estimation in a clinical setup [251].

Hu et al. devised an automatic segmentation framework based on 3D CNN and globally optimized surface evolution to overcome the difficulties of detecting and delineating liver from CT scans. A deep 3D CNN, inspired by the paper of Lu et al. [251], is trained to learn a subject-specific probability map of the liver, giving an initial surface and acting as a shape prior to the subsequent delineation phase. Then, a segmentation model adaptively incorporates global and local appearance information by performing a global optimization in a surface evolution fashion. The method has been validated on 42 CT scans coming from both the SLIVER07 dataset and local hospitals [252].

Kim et al. addressed the task of multi-organ segmentation of liver, duodenum, kidneys, and stomach with a 3D patch-based U-Net, trained with a standard six classes cross-entropy loss function. The authors concluded that the proposed method presents comparable results with inter-observer annotation of organs and outperforms atlas-based methodologies. Graph-cut post-processing proved to be useful as refinement in the segmentation setup [253].

Altini et al. developed a liver parenchyma and vessels segmentation method from CT scans in a surgical planning setup. The authors exploited 2.5D V-Net and compared different variants of Tversky Loss, exploiting ADAM optimizer [256]. The authors found a very

beneficial effect on hepatic vessels segmentation, resulting in an increase of the *Recall* index. In general, Tversky Loss should be considered for cases in which there are much more “negative” examples than “positive” ones, such as precise segmentation of small anatomical parts. Morphological post-processing helped the authors with the liver segmentation task [56].

2.2.3.3 Adversarial Loss setups

Xia et al. proposed an improvement over a basic framework with DeepLab-v3, introducing the pix2pix network as the GAN. The proposed model allows combining deep features with multi-scale semantic features for realizing the segmentation. The authors exploited a combination of three components in the loss formulation: segmentation-based loss (which is the standard multi-class cross-entropy loss), content-based loss (the function of the generator output), and adversarial-based loss (the function of the discriminator output). The proposed adversarial framework manages to overcome the space inconsistencies problems [254].

Yang et al. proposed an automatic Liver Segmentation setup exploiting an adversarial Deep Image-To-Image Network (DI2IN). They proposed a method trained on an annotated dataset containing 1000 CT volumes with variations in scanning protocols and patients conditions. DI2IN is designed as a symmetric convolutional encoder-decoder. All the blocks are 3D convolutional layers and bilinear upscaling ones. The adversarial network provides an additional loss contribution for updating the generator parameters during the training process [255].

2.2.3.4 Summary

A qualitative comparison of the considered methodologies and approaches for liver segmentation is reported in Table 2.4, whereas a quantitative comparison is reported in Table 2.5.

Significant differences can be observed among considered approaches due to datasets, sample sizes, kind of architectures, exploitation of 2D images or 3D ones, and assessment measures. Sophisticated 2D architectures (or, even better, 2.5D ones) can provide performances similar to 3D ones. The most used architectures are U-Net or its variations, as 2.5D V-Net [56], U-Net++ [88] or H-DenseUNet [87].

There is an amount of publicly available datasets which can be exploited to train and validate models for liver segmentation. In fact, publicly available datasets are used, at least partially, in almost all the papers dealing with the segmentation of liver images. The most

used datasets are SLIVER07 [126], 3D-IRCADb and LiTS [127]. Thanks to the LiTS dataset, it is now possible to have more than one hundred CT images with annotations of liver and lesions. It is worth noting that top-performing approaches include LiTS as a training set. With the exception of Taleb et al. [115], the other considered works only analyzed CT images for the task of liver segmentation.

Dice coefficient is the most widespread similarity measure for assessing performances of semantic segmentation of liver images. Moreover, VOE, RVD, ASSD, MSSD, which are among the quality measures for the SLIVER07 [126] and LiTS [127] challenges, also have become widely adopted.

Different authors found advanced post-processing techniques, like CRF, to have a beneficial effect on the segmentation performances [147, 249, 250]. Adversarial training applied to 2D images can allow obtaining performances on par with 3D architectures, for instance by exploiting pix2pix [254]. Moreover, adversarial approaches mitigate problems such as holes in the segmentation, since the discriminator can recognize this kind of artifact during the training process. SSL can enable to learn rich semantic feature representation that can be exploited to perform an accurate liver segmentation [115, 116].

2.2.4 Liver Segments Classification for Surgical Planning

There are some approaches proposed in the literature for liver segments' classification. Most of the published research tries to obtain the segments from an accurate modeling of the liver vessels.

Oliveira et al. employed a geometric fit algorithm, based on the least-squares method, to generate planes for separating the liver into its segments [239].

Yang et al. exploited a semi-automatic approach for calculating the segments [241]. The user has to input the root points for the branches of the portal vein and the hepatic veins, and then, a Nearest Neighbor Approximation (NNA) algorithm is implemented to assign each liver voxel to a specific segment. The classification was very straightforward: the voxel belonged to the segment whose branch was nearest to that voxel.

Among the approaches that aim to directly classify liver segments, it is worth noting the work of Tian et al. [210], who offered the first publicly available dataset for liver segments' classification. They realized an architecture based on U-Net [72], but with considerations for the Global and Local Context (GLC), resulting in a model that they referred to as GLC U-Net.

Table 2.4 Liver Segmentation with Deep Learning: a comparison.

Authors	Anatomical Sites	Dataset	Image Modality	DL Architecture	Post-processing	Adversarial Training	Evaluation Metrics
Zhou <i>et al.</i> [85, 88]	Liver	LiTS [127]	CT	U-Net++	No	No	IoU, Dice, Recall Specificity, F2 score
Li <i>et al.</i> [87]	Liver, Liver Tumors	3D-IRCADb, LiTS [127]	CT	H-DenseUNet	No	No	Dice, VOE, RVD, ASSD, MSSD
Ben-Cohen <i>et al.</i> [247, 248]	Liver, Liver Tumors	SLIVER07 [126], Local Dataset	CT	2.5D FCN	No	No	Dice, Recall, Precision
Christ <i>et al.</i> [147]	Liver, Liver Tumors	3D-IRCADb	CT	Cascaded FCNs	3D CRF	No	Dice, VOE, RVD, ASSD, MSSD
Dou <i>et al.</i> [249]	Liver	SLIVER07 [126]	CT	3D DSN	CRF	No	VOE, RVD, ASSD, RMSD, MSSD
Rafiei <i>et al.</i> [250]	Liver	BTCV [221, 222]	CT	3D-2D-FCN	CRF	No	Dice
Lu <i>et al.</i> [251]	Liver	SLIVER07 [126], 3D-IRCADb, Local Dataset	CT	3D CNN	Graph Cut	No	VOE, RVD, ASSD, RMSD, MSSD
Hu <i>et al.</i> [252]	Liver	SLIVER07 [126], Local Dataset	CT	3D CNN	Globally Optimization Surface Evolution	No	VOE, RVD, ASSD, RMSD, MSSD, Dice
Kim <i>et al.</i> [253]	Liver, Duodenum, Kidneys, Stomach	Local Dataset	CT	3D U-Net	Graph Cut	No	Dice, ASSD, MSSD
Altini <i>et al.</i> [56]	Liver, Liver Vessels	LiTS [127], SLIVER07 [126], 3D-IRCADb, Local Dataset	CT	2.5D V-Net (Tversky loss)	Morphological Operators	No	Dice, VOE, RVD, ASSD, MSSD, Accuracy, Recall, Specificity
Xia <i>et al.</i> [254]	Liver	LiTS [127], Local Dataset	CT	DeepLab-v3 + pix2pix	No	Yes	Dice, VOE, RVD, ASSD, MSSD, Accuracy, IoU
Yang <i>et al.</i> [255]	Liver	Local Dataset	CT	D12IN	No	Yes	Dice, ASSD
Taleb <i>et al.</i> [115]	Liver, Brain, Prostate	CHAOS [216–218]	CT + MRI	Multi-modal with SSL	No	Yes	Dice
Chung <i>et al.</i> [116]	Liver	SLIVER07 [126], 3D-IRCADb, Pancreas-CT [162, 209, 219], BTCV [221, 222], Local Dataset	CT	3D FCN with SSL	No	No	Dice, ASSD, MSSD, Recall, Precision

Table 2.5 *Performance comparison for Liver Segmentation*. * Note: for Li et al. [87], additional metrics are available only for the 3D-IRCADb dataset, but the authors used LiTS [127] as an additional train set, so they may have some CT scans in both train and test sets.

Authors	Dice Coefficient	Other Metrics	Sample Size	Input Size
Zhou et al. [85]	N/A	IoU = 0.83	331 CT images	2D: 512×512
Zhou et al. [88]	0.89 ± 0.03	IoU = 0.83 ± 0.01 , Recall = 0.93 ± 0.02 , Specificity = 0.99 ± 0.00 , F2 score = 0.91 ± 0.02	331 CT images	2D: 96×96
Li et al. [87]	0.96	N/A*	131 CT scans (train) + 70 CT scans (test)	2D DenseUNet + 3D aggregation
Ben-Cohen et al. [247]	0.89	Recall = 0.86, Precision = 0.95	20 + 20 patients	2D
Christ et al. [147]	0.94	VOE = 0.11, RVD = -0.01, ASSD = 1.5 mm, MSSD = 24.0 mm	15 CT scans	2D CNN + 3D CRF
Dou et al. [249]	N/A	VOE = 0.05 ± 0.01 , RVD = 0.01 ± 0.01 , ASSD = 0.67 ± 0.12 mm, RMSD = 1.48 ± 0.21 mm, MSSD = 29.63 ± 16.31 mm	30 CT images	3D: $160 \times 160 \times 72$
Rafiei et al. [250]	0.94	N/A	30 CT images	3D: $512 \times 512 \times 38$ (3D to 2D CNN)
Lu et al. [251]	N/A	SLIVER07: VOE = 0.06, RVD = 0.03, ASSD = 0.91 mm, RMSD = 1.88 mm, MSSD = 18.94 mm; 3D-IRCADb-01: VOE = 0.09 ± 0.03 , RVD = 0.01 ± 0.03 , ASSD = 1.89 ± 1.08 mm, RMSD = 4.15 ± 3.16 mm, MSSD = 33.14 ± 16.36 mm	30 CT images + 20 CT images	3D: $249 \times 249 \times 279$
Hu et al. [252]	SLIVER07-test (10 CT scans): 0.97 ± 0.01 , SLIVER07-train (20 CT scans): 0.97 ± 0.01 , Local (12 CT scans): 0.97 ± 0.00	SLIVER07-test (10 CT scans): VOE = 0.05 ± 0.01 , RVD = 0.00 ± 0.01 , ASSD = 0.84 ± 0.25 mm, RMSD = 1.78 ± 0.56 mm, MMSSD = 19.58 ± 3.07 mm; SLIVER07-train (20 CT scans): VOE = 0.05 ± 0.01 , RVD = 0.00 ± 0.02 , ASSD = 0.96 ± 0.24 mm, RMSD = 1.84 ± 0.54 mm, MSSD = 19.20 ± 5.97 mm; Local (12 CT scans): VOE = 0.05 ± 0.01 , RVD = 0.00 ± 0.02 , ASSD = 0.87 ± 0.14 mm, RMSD = 1.50 ± 0.31 mm, MSSD = 16.45 ± 6.47 mm	109 CT scans (train) + 10 CT scans (test) + 20 CT scans (test) + 12 CT scans (test)	3D: $496 \times 496 \times 279$
Kim et al. [253]	0.96 ± 0.02	MSSD = 0.714 ± 0.301 mm, ASSD = 8.926 ± 6.298 mm	80 CT scans (train) + 20 CT scans (validation) + 20 CT scans (test)	3D: $64 \times 64 \times 64$
Altini et al. [56]	0.97 ± 0.02	VOE = 0.07 ± 0.03 , RVD = 0.04 ± 0.04 , ASSD = 1.44 ± 0.89 mm, MSSD = 31.50 ± 12.05 mm	131 CT scans (train + validation) + 20 CT scans (test)	2.5D: $512 \times 512 \times 5$
Xia et al. [254]	0.97	VOE = 0.08, RVD = 0.01, ASSD = 1.925 mm, MSSD = 35.583 mm	12,150 CT slices (train) + 8,800 CT slices (test)	2D: 320×240
Yang et al. [255]	0.95 ± 0.02	ASSD = 1.90 ± 0.74 mm	>1000 CT scans	3D
Taleb et al. [115]	0.95	N/A	20 multimodal cases (train) + 20 multimodal cases (test)	2D
Chung et al. [116]	0.96 ± 0.01	ASSD = 1.17 ± 0.30 mm, MSSD = 5.25 ± 2.70 mm, Recall = 0.96 ± 0.03 , Precision = 0.96 ± 0.02	160 CT scans	3D: $128 \times 128 \times 64$

Yan et al. collected a huge CT dataset, composed of 500 cases, to realize CAD systems targeted at the liver region [257]. The ComLiver datasets include labels for liver parenchyma, vessels, and Couinaud segments. For the task of performing Couinaud segmentation, the authors compared only 3D CNN models, namely FCN [75], U-Net [72, 73], U-Net++ [85], nnU-Net [258], Attention U-Net (AU-Net) [259], and Parallel Reverse Attention Network (PraNet) [260].

2.2.5 Vertebrae Segmentation and Identification

The spine plays a primary role in sustaining and supporting the human body and shielding organ structures while allowing the full body mobility. It also protects the spinal cord from injuries and mechanical shocks due to impacts [261]. The anatomic complexity of the spine, which consists of 33 vertebrae, 23 intervertebral disks, the spinal cord and connecting ribs, often leads to an under-diagnosis of spinal pathologies [262]. The spinal surgeon is faced with the need of robust algorithms to segment and create a spine model, leading to the development of Computer-Assisted Surgery (CAS) systems [263]. The knowledge of the shape of single vertebrae can aid early diagnosis of degenerative disorders, spinal deformities or trauma and support surgical planning [264]. CT is the most spatially accurate imaging modality to assess the three-dimensional morphology of the vertebra [265].

The most significant challenges in the context of vertebrae segmentation and identification, including large-scale vertebrae segmentation challenges (VerSe‘19 and VerSe‘20), have been organized during the MICCAI international conferences [266, 267]. The data of VerSe‘19 are composed of 160 CT scans, whereas that of VerSe‘20 comprise 300 CT scans. Previously available datasets from other challenges in the spine imaging domain are much smaller. Examples include the Computational Spine Imaging 2014 Workshop, targeted at the segmentation of the thoraco-lumbar spine, which consists of 20 images [265, 268], and the online challenge xVertSeg, targeted at the segmentation of the lumbar spine, composed of 25 samples [264].

When elaborating spine imaging data, vertebrae classification and vertebrae segmentation are two pivotal tasks. Applications span from diagnosis (detecting and grading of vertebral fractures, spinal curve estimation, identification of spinal deformities), to biomechanical modeling and surgical planning for metal insertions. As a radiological imaging technique, CT scans are the gold standard for assessing the ‘bone’ part of the spine, since they guarantee high bone-to-soft-tissue contrast [261]. Several methods have been proposed in the literature for vertebrae segmentation and labeling. Traditionally, spine segmentation has

been approached predominantly as a model-fitting problem; however, more recent spine segmentation techniques focused on DL-based methods [269].

2.2.5.1 Related Work

Kim et al. developed a web-based tool for spine imaging data segmentation from CT scans [270], exploiting deep learning-based methodologies, especially the U-Net architecture [72]. The tool was implemented in Python with the Keras library for the data processing side, whereas a Flask server framework was developed for providing accessibility over the web. The U-Net was trained on 310 images from CT scans, validated on 20 images and tested on only 14 images. This approach allowed the authors to obtain a Dice coefficient of roughly 90% for the binary spine segmentation task.

Vania et al. proposed a method for the automatic spine segmentation from CT scans using CNN via the generation of redundant class labels [263]. The implemented architecture consisted of two convolutional layers and three fully connected ones. Besides classes for background and spine, two redundant classes were generated by dilating the spine mask. This approach allowed the authors to obtain a Dice coefficient of 94% for the binary spine segmentation task.

Qadri et al. developed an automatic approach, named patch-based deep belief networks (PaDBNs), for vertebrae segmentation in CT images [271]. Deep belief networks (DBNs) are DL models composed of stacked Restricted Boltzmann Machines (RBMs) [272]. Their proposed model allowed them to automatically select the features from image patches and measure the difference between classes. Unsupervised learning was exploited for weight initialization, whereas supervised fine-tuning was used to update weights. One strength offered by this methodology is the considerable reduction in the computational cost while retaining good performances.

Zareie et al. [273] proposed two methods for vertebrae segmentation in 3D CT images. The first is based on a multi-layer perceptron neural network (MLPNN), which used seven gray-level statistical features to classify each voxel as vertebrae or background. The second method implemented an Adaptive Pulse Coupled Neural Network (APCNN) to segment vertebrae and used a median filter to refine the results. This network was a modified version of the one proposed by Chang et al. [274], in which the parameters were adjusted adaptively according to the input image. The performances of both systems were calculated in terms of *DSC* (definition is provided in Section 2.1.7) on seventeen 3D vertebrae CT images of the thoracic and lumbar spine of both normal and abnormal cases. The results compared four different models: the PCNN developed by Chang et al. [274], the APCNN, the MLPNN with

all seven features and the MLPNN using only the intensity level of each voxel as a feature (MLPNN1f). In addition, the robustness of APCNN and MLPNN was evaluated by adding salt-and-pepper noise to the images. It was shown that the APCNN performed better than the other methods with a *DSC* of 95%, being less sensitive to noise than the MLPNN and more adjustable to each image than a classic PCNN.

All the works considered so far did not address the recognition of the different vertebrae, which is a considerably more complicated task than the binary spine segmentation. Slightly more sophisticated approaches can allow vertebrae identification as a post-processing step after the spine segmentation stage, using simple and not-trainable techniques, which do not require apposite dataset preparation.

Bae et al. proposed a fully automated approach for 3D segmentation and separation of multiple cervical vertebrae in CT images exploiting a 2D CNN [275]. The authors trained a 2D U-Net model for performing the spine segmentation, considering two classes for the superior and the inferior part of the vertebra, obtaining accuracies comparable with the inter- and intra-observer variability of the manual segmentation performed by human experts. In order to separate the vertebrae, the authors proposed a post-processing stage. In the first part, each class region continuity is assessed by using the connected component analysis to correct mis-labeling errors. Then, a technique for detecting separation points across the superior and inferior region was implemented. After having identified these points, voxels belonging to superior and inferior part of the same vertebrae were merged, leading to a final segmentation with distinct vertebrae.

In the realm of the methods proposed for the VerSe challenges, it is worth considering the top three works from the VerSe'19 challenge [261, 267, 269, 276] for their accurate results. All the methods that have been proposed for these challenges also considered the vertebrae identification problem.

Sekuboyina et al. [267] proposed a pipeline to localize and identify the vertebrae on multidetector CT scans employing an improved version of the Btrfly Net [277]. The Btrfly fully convolutional neural network works on sagittal and coronal 2D projections and incorporates the spine localization and anatomic prior information using the adversarial discriminators. The approach was tested on both a public dataset of 302 CT scans and two in-house datasets with a total of 238 CT scans. On the public dataset, the network achieved a vertebrae identification rate of 88.5%. On the in-house datasets, instead, with a higher interscan data variability, an identification rate of 85.1% was obtained. One of the principal limitations of the study was that it only considered 24 labels for C1-L5 and did not account

for segmentation anomalies, such as L6, or transitional vertebrae, such as a lumbarized S1 vertebra.

Lessman et al. proposed iterative FCNs. The idea was that an instance memory keeps information about previously segmented vertebrae. This memory was then combined with an FCN. This network iteratively analyzed image patches, searched for the first not yet segmented vertebra, which was recognized as completely or partially visible, so that partially visible vertebrae were excluded from further analysis [269].

Payer et al. proposed a cascaded approach which involves three stages. In the first step, a U-Net [72] variation was exploited to regress a heatmap of the spine centerline, which was generated by combining Gaussian heatmaps of all the individual landmarks; this allowed to locate the approximate position of the spine. In the second phase, SpatialConfiguration-Net was employed to localize centers of the vertebrae bodies. It effectively combines local aspect of reference points with their spatial configuration. In the last stage, a U-Net trained with cross-entropy was exploited for the binary segmentation of single separated vertebrae [276]. A more detailed list of approaches considered for the VerSe challenges is presented in [261].

2.2.6 Prostate Fusion Biopsy

Prostate cancer is a major health problem and represents the most common cancer in the male population, accounting for 18.5% of all the cancers diagnosed in humans [278]. The number of new cases worldwide crossed 1,275,000 and caused approximately 360,000 deaths merely in year 2018 (3.8% of all deaths caused by cancer in men) [279].

Numerous imaging modalities are exploited for prostate cancer diagnosis, treatment and follow-up. The Transrectal Ultrasound (TRUS), MRI, and CT are the most common employed imaging modalities [280]. Each technique provides different information and is used for several divergent clinical scopes. During biopsy procedures, TRUS is commonly employed since it is an inexpensive, portable and real-time methodology [281]. MRI is mainly adopted for diagnosis and treatment planning [282]. In fact, this modality has a better soft tissue contrast and allows a more efficient lesion detection and staging in patients affected by prostate cancer.

As can be seen from Figure 2.6, the TRUS images suffer from problems as speckle, low contrast and shadow artifacts [283]. Calcification and acoustic shadowing make the automatic segmentation of prostate region a very complex task [284]. The prostate usually appears like a hypoechoic mass encompassed by a hyperechoic region [285]. CT scans are useful in

determining if prostate cancer has spread to bone tissues or to assess the effectiveness of the brachytherapy [286].

In the clinical practice, majority of prostate cancer cases are diagnosed, prior to symptoms development, thanks to the prostate-specific antigen (PSA) [287] levels in the blood and rectal examination. In order to achieve more satisfactory information, MRI is the election modality, with PI-RADS v2.1 being the standard for finding interpretation [288].

The standard prostate biopsy involves the extraction of 10 to 12 tissue samples. Since there is no guarantee that sampling prostate in these regions is the most effective way to obtain the regions with cancerous tissue, fusion guided prostate biopsy is becoming now the preferred modality for most urologists and surgeons. In this way, suspicious areas found in the MRI of the prostate can be targeted during the prostate procedure exploiting the fusion with the real-time TRUS imaging, also allowing a better view of the biopsy needle. Advantages of this approach comprises of the following: more accurate sampling of the cancerous tissue in the prostate gland; less amount of patient tissue is extracted; less pain and less risks for the patient, including faster recovery time [289].

In order to implement a fusion prostate biopsy framework, segmentation of prostate gland must be obtained from both TRUS and MRI domains. Exceptions involve systems in which images are manually registered by the user at procedure time, by superimposing MRI over the TRUS. Since MRI is acquired days before the prostate biopsy, it is not fundamental that its segmentation is performed real time. Even though, manual segmentation of prostate from MRI is a tedious task and prone to inter- and intra-radiologist variation [290], so the exploitation of an automatic method grounded on the nnU-Net framework [258] for this task can further ease the procedure and improve its diagnostic accuracy. It is worth noting that nnU-Net does not denote a novel network topology, loss function, or training procedure. Indeed, nnU-Net stands for “no new net”. The strength of the nnU-Net framework comes from the systematization of all the steps which were usually manually tuned in the training pipeline of semantic segmentation architectures, including data augmentation, hyperparameters’ tuning, test-time augmentation, and ensembling. Instead, manual segmentation of the prostate gland from TRUS has to be realized real-time during the prostate procedure, therefore the need for a fast and effective methodology for this task is really fundamental in the clinical practice.

Ghose et al. performed a comprehensive survey which focused on methods for prostate segmentation in TRUS, MR, and CT images [280]. Prostate gland segmentation eases multi-modal image fusion for tumor localization in biopsy. Manual annotation of radiological images is a tedious and error prone task, which also has problems like inter- and intra-radiologist variability. Fully automatic methods, as those based on DL, require huge annotated

data, usually of the same transducer that will be used for the procedures, since there is a high variability in ultrasound image quality across vendors. Nonetheless, when data is available, DL methodologies show their strength, as is the case for Deep Attentional Features (DAF) [291] and DAF 3D [292]. The shortcoming of these techniques is that they cannot be applied before having acquired a dataset with images of the same ultrasound device that will be adopted during procedures.

Mahdavi et al. [293] proposed a semi-automatic prostate segmentation method that can be applied in prostate brachytherapy setups. The 3D geometric model of the prostate is created based on prior knowledge of the shape of the gland and on the assumption that the prostate has a tapered ellipsoidal shape and is slightly warped posteriorly due to the presence of the TRUS probe. They used, as prior shape of prostate gland, a tapered and warped ellipsoid. The proposed segmentation algorithm requires a manual initialization of the physician: on the mid-gland image the user selects six boundary points following a specific criterion. The main disadvantage of this method is that it requires the user to put initialization points in a very precise way, and relies deeply on these points, posing problems if they are slightly inaccurately placed or when the prostate region has an irregular shape.

Gong et al. [294] incorporated deformable superellipses in a Bayesian segmentation framework, exploiting an edge detection algorithm for discovering prostate boundaries. They show the capacity of deformable superellipses to capture the prostate shape in various anatomical zones. The main limitation of this method is that it requires to have an initial contour that is similar to the real boundaries of the prostate gland. To overcome this issue, Saroul et al. [295] proposed a variational approach, exploiting the implicit representation of a superellipse for modeling the active contour.

2.2.7 Lung COVID-19 Characterization

Since the beginning of the COVID-19 pandemic, researchers have focused on the association between interstitial pneumonia related to the infection and onset of thromboembolic phenomena. In fact, several researchers have found that the clinical aggravations of some patients are correlated to an increase in some laboratory indices, such as D-dimers and fibrinogen, and to the finding of pulmonary embolism (PE) or pulmonary thromboembolism (PTE) detectable with a dedicated computed tomography angiography (CTA) study [296–298]. Therefore, the onset of PTE in COVID-19 patients is considered a prognostically unfavorable event that requires prompt treatment with low molecular weight heparin or other anticoagulants [299]. Other researchers have proposed elaborations based on neural topologies in case series of

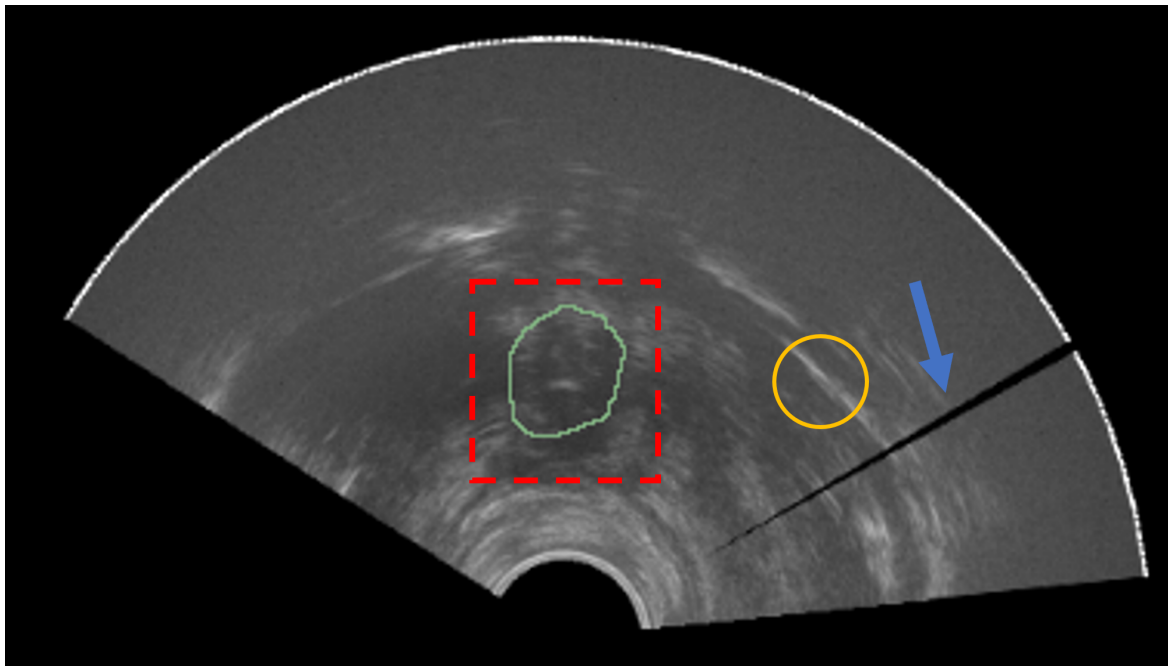


Fig. 2.6 *Prostate gland in TRUS image*. Prostate apex (ground truth mask in green) is not well distinguishable from the rest of the image (red dashed box). Yellow circle represents an example of region with low signal-to-noise ratio. Blue arrow denotes a shadow artifact.

COVID-19 patients in an attempt to provide alternative diagnostic screening tools to the nasopharyngeal swab. The latter still represents the diagnostic gold standard, even if it is expensive, with long sample processing times [300].

Prior to the start of the pandemic, the possibility to detect PTE from CT scans was already considered in some studies [301–306].

Cobelli et al. performed a comparison between the CT with and without contrast medium. Exploiting the supervision of an expert radiologist, the authors concluded that in 21 out of 51 examined cases, it was possible to detect the embolus in the unenhanced scan [301].

Mohamed et al. compared the diagnosis of pulmonary embolism from unenhanced CT and CTA scans, by determining the sensitivity and specificity of direct and indirect signs from the latter ones [303].

Sun et al. evaluated the detection of pulmonary embolisms on non-contrast CTs, which were reviewed by two radiologists. Subsequently, a third radiologist performed a final validation. The authors concluded that the non-contrast CTs were not sensitive enough to detect the PE [304].

Tourassi et al. created a CAD system, based on ANNs, to predict PE from ventilation-perfusion scans and chest radiographs. Their model outperformed physicians' predictions

[305]. Thus, the CTA scan is the election method for diagnosing PTE. Some researchers opted to employ CT scans, but the results were not always satisfactory.

Several recent studies addressed the task of lung parenchyma and lesion segmentation from CT scans in COVID-19 patients [307–317].

Wu et al. proposed a joint classification and segmentation (JCS) system in order to address the issue of performing real-time and explainable COVID-19 diagnosis from CT scans [307]. Akbari et al. developed a method for COVID-19 lesion segmentation from CT images based on active contour models [308].

Cao et al. exploited a CNN based on the U-Net architecture for illustrating the potential of deep learning-based quantitative CT scans for providing an objective assessment of pulmonary involvement and therapy response in COVID-19 [309].

Rajinikanth et al. developed a method that relies on the firefly algorithm and multi-thresholding based on Shannon entropy to enhance the pneumonia lesion. Finally, the Markov random field segmentation is exploited to extract the lesions with higher accuracy. According to the authors, the proposed workflow permitted to obtain a mean accuracy higher than 92% for COVID-19 lesion segmentation [310]. Rajinikanth et al. also proposed an image-assisted system to extract COVID-19 infected sections from lung CT scans in coronal view. The approach considered by the authors incorporates the following steps: thresholding to extract the lung region; image enhancement exploiting harmony search optimization and Otsu thresholding; image segmentation and ROI extraction from the binary image to compute the level of severity [311].

Ter-Sarkisov et al. proposed a one-shot model by exploiting Mask R-CNN [93] functionalities for the COVID-19 classification and lesion segmentation in chest CT scans utilizing the affinity between lesion mask features [312].

Zhao et al. built an open-source dataset composed of 349 COVID-19 CT images, coming from 216 patients, and 463 non-COVID-19 CT images. They also performed experimental studies for developing diagnostic methods based on multi-task learning and self-supervised learning, achieving an F_1 -score of 90%, an AUC-ROC (Area Under the Curve – Receiver Operating Characteristic) of 98% and an accuracy of 89% [313].

Wang et al. presented their experience in building and deploying an AI system that automatically analyses CT images and provides the probability of infection to rapidly detect COVID-19 pneumonia. By exploiting 1136 training samples (723 were positive for COVID-19) from five hospitals, the authors were able to achieve sensitivity of 97.4% and specificity of 92.2% on the test dataset [314].

Oulefki et al. realized an automatic COVID-19 lung infected region segmentation and measurement approach from CT scan images. The automatic tool proposed by the authors, compared to the state-of-the-art segmentation approaches, i.e., GraphCut, Medical Image Segmentation (MIS), and Watershed, showed better results with an accuracy of 98% [315].

Zheng, Wang et al. proposed a deep learning-based method for COVID-19 detection from chest CT scans, exploiting weak annotations. The lung region was segmented with a pre-trained U-Net. Then the segmented 3D lung region was fed into a 3D deep neural network to predict the probability of COVID-19 infection. The algorithm allowed to obtain an accuracy of 90.1% [316, 317].

2.2.7.1 Radiomic Features Definitions

This section introduces definitions for the radiomic features considered for the research carried out for this thesis. The interested reader may consult PyRadiomics documentation [318] or the IBSI reference manual [319] for further information.

The gray-level co-occurrence matrix (GLCM) describes the second-order joint probability function of an image ROI. The $(i, j)^{th}$ element of this matrix represents the occurrences of levels i and j separated by δ voxels along the θ direction in the image; that is $P_{GLCM}(i, j | \delta, \theta)$. PyRadiomics by default computes the symmetric GLCM, resulting in a symmetric P matrix. The normalized version can be obtained as in Equation (2.28).

$$p_{GLCM}(i, j) = \frac{P_{GLCM}(i, j)}{\sum P_{GLCM}(i, j)} \quad (2.28)$$

Then, the GLCM maximum probability (MP_{GLCM}) can be defined as in Equation (2.29).

$$MP_{GLCM} = \max p_{GLCM}(i, j) \quad (2.29)$$

In order to define the GLCM inverse variance (IV_{GLCM}), the diagonal probabilities need to be introduced, as in Equation (2.30):

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{GLCM}(i, j) \quad (2.30)$$

with $|i - j| = k; k = 0, 1, \dots, N_g - 1; N_g$ represents the number of discrete intensity levels in the image. Then, IV_{GLCM} can be defined as in Equation (2.31):

$$IV_{GLCM} = \sum_{k=1}^{N_g-1} \frac{p_{x-y}(k)}{k^2}. \quad (2.31)$$

The GLCM cluster shade (CS_{GLCM}) measures the skewness and uniformity of the GLCM, and can be defined as in Equation (2.32):

$$CS_{GLCM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{GLCM}(i, j) \cdot (i + j - \mu_x - \mu_y)^3 \quad (2.32)$$

where μ_x and μ_y can be defined as in Equation (2.33) and Equation (2.34), respectively.

$$\mu_x = \sum_{i=1}^{N_g} i \cdot p_x(i) \quad (2.33)$$

$$\mu_y = \sum_{j=1}^{N_g} j \cdot p_y(j) \quad (2.34)$$

Lastly, $p_x(i)$ and $p_y(j)$ can be represented as in Equation (2.35) and Equation (2.36), respectively.

$$p_x(i) = \sum_{j=1}^{N_g} P_{GLCM}(i, j) \quad (2.35)$$

$$p_y(j) = \sum_{i=1}^{N_g} P_{GLCM}(i, j) \quad (2.36)$$

The gray-level size zone matrix (GLSZM) measures the number of zones of the connected voxels that have the same gray-level intensity. The $(i, j)^{th}$ element represents how many zones with gray-level i and size j appear in the image. The GLSZM will be denoted as $P_{GLSZM}(i, j)$. The normalized GLSZM, $p_{GLSZM}(i, j)$, can be obtained as in Equation (2.37):

$$p_{GLSZM}(i, j) = \frac{P_{GLSZM}(i, j)}{N_z} \quad (2.37)$$

where N_z is the number of zones in the ROI.

The GLSZM gray-level variance (GLV_{GLSZM}) measures the variance in gray-level intensities for the zones, as shown in Equation (2.38):

$$GLV_{GLSZM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} p_{GLSZM}(i, j) \cdot (i - \mu_{GLSZM})^2 \quad (2.38)$$

where N_s is the number of discrete zone sizes in the image and μ_{GLSZM} can be defined as in Equation (2.39):

$$\mu_{GLSZM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} i \cdot p_{GLSZM}(i, j) \quad (2.39)$$

Gray-level dependence matrix (GLDM) portrays the gray-level dependencies in an image. They are defined as the number of connected voxels inward a distance δ from the center voxel. If $|i - j| \leq \alpha$, where α is a customizable cutoff parameter, the neighboring voxel with gray-level j is regarded as dependent on the center voxel having gray-level i . The $(i, j)^{th}$ element describes the number of times a voxel with gray-level i with j dependent voxels in its neighborhood appears in the image. The GLDM will be denoted as $P_{GLDM}(i, j)$. The normalized GLDM, $p_{GLDM}(i, j)$, can be obtained as in Equation (2.40):

$$p_{GLDM}(i, j) = \frac{P_{GLDM}(i, j)}{N_{dz}} \quad (2.40)$$

where N_{dz} is the number of dependency zones in the ROI.

The GLDM dependence entropy (DE_{GLDM}) can be defined as in Equation (2.41):

$$DE_{GLDM} = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} p_{GLDM}(i, j) \cdot \log_2(p_{GLDM}(i, j) + \varepsilon) \quad (2.41)$$

where N_d is the number of discrete dependency sizes in the image and ε is a small constant for avoiding numeric errors.

Gray-level run length matrix (GLRLM) analyzes gray-level runs, namely the number of consecutive voxels having the same gray-level value. In a GLRLM, $(i, j)^{th}$ element represents, along a given direction θ , the number of runs with gray-level i and length j which occur in the image. The GLRLM will be denoted as $P_{GLRLM}(i, j | \theta)$. The normalized GLRLM, $p_{GLRLM}(i, j | \theta)$, can be obtained as in Equation (2.42):

$$p_{GLRLM}(i, j | \theta) = \frac{P_{GLRLM}(i, j | \theta)}{N_r(\theta)} \quad (2.42)$$

where $N_r(\theta)$ is the number of runs in the image along the direction θ .

GLRLM gray-level variance (GLV_{GLRLM}) assesses the variance in gray-level intensities for the runs, as defined in Equation (2.43):

$$GLV_{GLRLM} = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p_{GLRLM}(i, j | \theta) \cdot (i - \mu_{GLRLM})^2 \quad (2.43)$$

where N_r is the number of discrete run length in the image and μ_{GLRLM} is defined as in Equation (2.44).

$$\mu_{GLRLM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} i \cdot p_{GLRLM}(i, j | \theta) \quad (2.44)$$

The long-run low gray-level emphasis (LRLGLE) measures the joint distribution of long-run length with low gray-level values, as defined in Equation (2.45):

$$LRLGLE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p_{GLRLM}(i, j | \theta) \cdot j^2}{i^2}}{N_r(\theta)} \quad (2.45)$$

Neighboring gray-tone difference matrix (NGTDM) evaluates the dissimilarity between a gray value and the average value of the neighboring voxels inward a distance δ . Inside the matrix, the absolute differences sum for gray-level i is stored. Let X_{gl} be a set of delineated voxels and $x_{gl}(j_x, j_y) \in X_{gl}$ be the gray-level of a voxel at location (j_x, j_y) , then the average gray-level of the neighborhood ($\bar{A}_i$) can be defined as in Equation (2.46):

$$\bar{A}_i = \bar{A}(j_x, j_y) = \frac{1}{W} \sum_{k_x=-\delta}^{\delta} \sum_{k_y=-\delta}^{\delta} x_{gl}(j_x + k_x, j_y + k_y) \quad (2.46)$$

where $(k_x, k_y) \neq (0, 0)$; $x_{gl}(j_x + k_x, j_y + k_y) \in X_{gl}$; W represents the number of voxels in the neighborhood.

The contrast (C_{NGDTM}) is a measure of spatial intensity variability, and can be defined as in Equation (2.47):

$$C_{NGDTM} = \left(\frac{1}{N_{g,p}(N_{g,p} - 1)} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_i \cdot p_j (i - j)^2 \right) \left(\frac{1}{N_{v,p}} \sum_{i=1}^{N_g} s_i \right) \quad (2.47)$$

with $p_i \neq 0$, $p_j \neq 0$; $N_{g,p}$ is the number of gray-levels where $p_i \neq 0$; $N_{v,p}$ is the number of voxels in X_{gl} ; $p_i, i = 1, \dots, N_g$ is the relative frequency of the i gray-level; $s_i = \sum |i - \bar{A}_i|$.

Inside the realm of first-order statistics, the skewness (g_1) measures the asymmetry of the distribution of values about the mean value, and can be defined as in Equation (2.48):

$$g_1 = \frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \hat{X})^3}{\left(\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \hat{X})^2 \right)^{3/2}} \quad (2.48)$$

where N_p is the number of voxels in the image.

2.3 Pathomics and Deep Learning in Digital Pathology

Digital Pathology is changing the way in which histopathological diagnosis is made. Indeed, in current clinical practice, Digital Pathology is progressively becoming a technological prerequisite for scientific laboratories [9]. The factors which are contributing to this change of paradigm include the availability of whole-slide imaging, fast computer networks and low-cost storage solutions, which are easing the management and sharing of WSIs for the pathologists [9]. Furthermore, the advent of Machine Learning and Deep Learning solutions for the automatic processing of histopathological tissue is offering possibility for image-based diagnosis, prognosis and treatment. Computer-aided Diagnosis and Clinical Decision Support Systems can have a pivotal role in assisting pathologists, extending their perspective with precise quantitative analysis.

An emerging paradigm which is linked to Digital Pathology is Pathomics. According to Niazi et al. [320], the term Pathomics refers to the possibility of incorporating the broad range of image analyses data to extract quantitative features that can characterize diverse phenotypic characteristics of tissue samples in WSIs. Image analysis workflows in Digital Pathology involve object detection, segmentation, labeling and classification of tissue, considering histological features such as architectural elements, nuclei, and cells. The traditional histopathological assessment carried out by pathologists can be accompanied by quantitative features which map phenotypic patterns. The fundamental aim of these analyses is to blend pathomics with clinical, radiological, and genomic data to obtain models capable to predict clinical outcomes and response to treatment in the most accurate way [320].

2.3.1 Stain Color Normalization in CRC Histopathology

CRC is the second cause of death for cancer with mortality reaching almost 35% [321]. In the last few years, new targeted therapies have been developed gaining significant improvement

in clinical outcomes for several malignancies. To date, a shift from tumor cells to the tumor microenvironment (e.g., for immunotherapeutic treatments) highlighted the importance to know cell-cell interaction in the context of tissue morphology. As an example, there is a growing interest in the knowledge of the spatial location of transcriptomic data. Thus, the segmentation of tissue types is required to better perform spatial analyses, such as for the selection of relevant regions of interest. Moreover, it is well-known that stroma-rich CRC has a poor prognosis [322], and a tissue segmentation pipeline could be helpful in prognosis prediction.

Image processing and DL techniques can be exploited for the automatic analysis of histological images, e.g., tissue type classification. The traditional workflow for realizing an image classifier is composed of several stages, i.e. preprocessing, feature extraction, dimensionality reduction, and classification, that can be obtained with models like SVMs, Decision Trees (DTs), and ANNs. DL-based workflows, instead, can enable end-to-end training of the models, easing complex steps, such as handcrafted feature extraction, and leading to performance improvement. These techniques can be exploited for developing CAD systems which can enhance pathologists' workflows, reducing issues concerning intra- and inter-pathologist variability [323].

Nevertheless, a fundamental problem regarding the histopathological classification of images arises from the differences in colors between tissue samples from different institutions. Indeed, a complex protocol composed of several steps, namely (i) collection and fixation, (ii) dehydration and clearing, (iii) paraffin embedding, (iv) microtomy, (v) staining, (vi) mounting, and (vii) digitalization, is required for Digital Pathology workflows [323]. Artifacts and differences among laboratories can be introduced at any of these stages [51].

Stain color normalization is therefore a pivotal pre-processing step for successfully deploying Deep Learning CAD frameworks in Digital Pathology setups [324, 325]. Currently, the taxonomy of stain color normalization methods comprises: (a) global color normalization; (b) color normalization after stain separation; (c) color transfer with deep neural networks [51]. The first two approaches involve traditional image processing techniques, whereas the third harnesses the power of DL. Indeed, in recent years, the possibilities offered by GANs, to effectively implement color transfer between histopathological images, are improving the capabilities of classification systems by creating higher quality normalized images [326]. Other authors considered the problem of stain color normalization in nuclei segmentation pipelines [327], or for normalization of tissue of breast and prostate cancers [328, 329], but most of these works limit their analysis on cGAN [330], CycleGAN [99] and its variants, as Residual Cycle-GAN [21].

The Image-to-Image Translation paradigm [330] can be thought as a general framework to tackle a variety of image analysis problems, such as segmentation, color normalization, reconstruction of original images from labels, and conversion from one source modality to another, among the others. This approach consists in training a conditional GAN for translating images from a source domain to a target domain; however, such architectures need a dataset with paired images for setup and training. On the other hand, CycleGAN and several subsequent works focused on the idea to offer this translation having only the domain as the label, without the need for paired data, leading to the concept of Unpaired Image-to-Image Translation (UI2IT). This paradigm allows to construct datasets for normalization in a manner that is affordable for Digital Pathology laboratories, as paired image data is normally not available, especially when data come from two distinct institutions.

2.3.1.1 Related Works

Colorectal Cancer Tissue Classification The issue of classifying epithelium and stroma from digitized tumor tissue microarrays (TMAs) has been considered by Linder et al. in 2012 [331]. In the feature extraction phase, the authors took advantage of LBP (Local Binary Patterns) and LBP/C, where C is a contrast measure. Other features considered were Gabor filtered images and Haralick texture features [332, 333]. Employing an SVM classifier, the authors stated that the LBP/C-based is the best one, with an AUC-ROC of 0.995.

Kather et al. considered a multi-class tissue classification in the domain of colorectal cancer histopathology [334]. During the feature extraction stage, they considered several categories of features, after having transformed the original color images into gray-scale ones: histogram features, of both lower-order and higher-order; LBP; Gabor filters; GLCM; perception-like features. For the classification step, the authors investigated four classifiers: decision trees, linear SVM, radial-basis function SVM, and 1-nearest neighbor. The same feature set can obtain higher results by exploiting red channel versions of images instead of gray-scale ones [20], even though this observation holds mainly for unnormalized images. Kather et al. also exploited the capabilities of CNNs for the sake of classifying CRC H&E histopathology images of The Cancer Genome Atlas (TCGA) composed of 862 WSIs [25].

From the literature analysis of works for CRC tissue classification, it emerges a lack of exploration of methodologies for stain color normalization. Indeed, most of these studies are tailored to discover the most efficient features or classification architecture for the task in hand, without a proper consideration of the pre-processing steps as stain color normalization.

Stain Color Normalization for Colon Histological Tissue Several DL-based methods have been proposed to tackle the stain color normalization problem for histopathological tissues including the colon [21, 327, 335–337].

Bentaieb et al. proposed a stain transfer based approach for stain color normalization [335]. The authors designed a discriminative image analysis network that has the capability to relocate stains between different datasets. Their architecture is composed of a generative network devoted to learning both dataset-specific staining properties and image-specific color transformation, and a task-specific network which is exploited for the downstream task (as segmentation or classification). Their model can be trained end-to-end by exploiting a multi-objective loss. The authors conclusion states that their model is capable of improving the results, both for what concerns the quality of normalized images, and the accuracy of the networks for the downstream tasks, over various baselines.

Pontalba et al. evaluated the impact of several existing methodologies for stain color normalization in a setup for nuclei segmentation [327]. The considered normalization techniques comprise histogram specification, color transfer, stain specific color transfer, spectral matching, and CycleGAN. The authors also considered several measures for assessing quality of normalized images, besides evaluating results of the downstream segmentation task. To enhance the CycleGAN capabilities, de Bel et al. proposed an improvement over the base model, with the embodiment of residual learning, devising an architecture that they define as Residual CycleGAN [21]. The authors compare the performance of their stain normalization approach, also with respect to data augmentation, to prove the robustness of the downstream segmentation networks. The considered downstream applications include segmentation from colon and kidney tissue samples.

Shen et al. noted that the transformation induced by GANs can cause information loss, or suffer from problems such as mode collapse, damaging results for the subsequent diagnostic task [336]. To solve this problem, they devised a contrastive learning method with a color-variation constraint, to retain the recognizable phenotypic features when using a GAN for stain color normalization. Self-supervised learning allows to cluster discriminative tissue patches between several types of tumors.

Kausar et al. introduced a deep model, which they defined as Stain Acclimation Generative Adversarial Network (SA-GAN), which has an architecture that comprises one generator and two discriminators [337]. As usual, the purpose of the generator is to transform images from the source domain to the target domain. The two discriminators, instead, have different roles: the first one enforces the generated images to retain the color patterns of the target

domain, whereas the second one enforces the generated images to maintain the structure contents of the source domain.

With respect to the work summarized for the stain color normalization, it is possible to observe that: (i) other authors have not applied these techniques for the application of multiclass CRC tissue classification; (ii) other authors did not always found a suitable way to include pathologists evaluations or to correlate them with existing metrics for assessing GAN-generated image quality or accuracy of the downstream task; (iii) other authors have not explored the whole UI2IT framework, but rather focused on single models, especially those similar to CycleGAN [21, 327] or contrastive learning [336].

2.3.2 Nuclei Detection from Histopathological Images

In the healthcare scenario, artificial intelligence is exploited in medical imaging as a powerful tool with which to characterize objects of interest and lesions in anatomical regions under consideration. Traditionally, pathologists manually analyze numerous biopsies or tissue samples to diagnose complex pathologies, such as cancer. Even though it is tedious and time-consuming, this approach remains the gold standard [338, 339].

Computational pathology attempts to overcome the main challenges arising from manual histological image evaluation, such as inter- and intraobserver variability or the inability to evaluate the smallest visual features and the time required to examine WSIs [23, 338, 340].

The nuclei of cells provide a large amount of information for the analysis of histopathological tissue. For instance, immunohistochemistry-marked nuclei can be exploited for the estimation of cellular proliferation in cancer (e.g., Ki-67). Hence, nuclei segmentation is a fundamental first step toward the automated analysis of WSIs [341]. However, the difficulties associated with variable coloring arising from H&E-stained images, overlapped nuclei, the presence of artifacts, and differences in cell morphology and texture, represent obstacles for computer-based segmentation algorithms [23, 339]. Moreover, WSIs have very high resolutions and contain an enormous number of nuclei, adding peculiarity to the task [342]. A critical aspect in several computational pathology pipelines is to achieve accurate segmentation of nuclei both for subsequent extraction and classification of nucleus features, but also for analyzing cellular distribution, useful for classifying tissue subtypes and identifying abnormalities [23].

Several studies focused on nuclei detection because of its importance in the pathologic diagnostic pipeline, in particular in the field of oncology. As an example, nuclei detection could be helpful to distinguish nuclei undergoing changes, indicating a progression of

squamous epithelium cervical intraepithelial neoplasia [343]. Moreover, the estimation of tumor cellularity is very important, particularly in the era of precision medicine. Indeed, bioinformatic pipelines for copy number variation analysis require tumor cellularity as input and for a correct evaluation of variant allelic frequency [344].

Machine Learning-based nuclear segmentation methods are typically the most efficient, as they can learn to identify variations in the shape and coloration of nuclei. In the semantic segmentation [33, 35] approach, all image pixels are labeled as nuclear or background through a deep learning model. Nevertheless, these methods often fail to distinguish the different instances of objects of interest, i.e., nuclei, which then need to be addressed with ad hoc post-processing techniques, such as clustering [18].

The detection task can be approached by exploiting morphological features. CRImage [345] profits from thresholding as the first step for nuclei detection. Centroids of segmented nuclei are used as the point of detection. Then, a list of statistics for each segmented nucleus is utilized as a feature vector, and classification involves a support vector machine with radial basis kernel. Finally, spatial density smoothing is used to correct false detections.

LIPSyM [346] introduces the local isotropic phase symmetry measurement, designed to give high values to cell centers and nearby pixels; on the other hand, it cannot precisely detect spindle-like and other irregularly shaped nuclei such as fibroblasts and malignant epithelial nuclei.

In the last several years, CNNs are emerging as the most effective way to tackle the nuclei detection task. In particular, the spatially constrained convolutional neural network (SC-CNN) [347] uses spatial regression for localizing the nuclei centers; the regression in SC-CNN is model-based, which explicitly constrains the output form of the network.

Xu et al. [342] used a stacked sparse autoencoder (SSAE) to learn a high-level representation of nuclear and non-nuclear objects by means of a softmax classifier.

Finally, the R2U-Net-based regression model named “UD-Net” [340] is proposed for end-to-end nuclei detection from pathological images. The recurrent convolutional operations help the model learn and represent features better than the feed-forward convolutional operations, and the robustness of the R2U-Net model has been demonstrated previously in several studies [348].

2.3.3 Computer-Aided Diagnosis for IgA-GN Oxford Classification

In the realm of nephropathology, a fundamental role is reserved to the assessment of glomerular structures and potential lesions. IgA Glomerulonephritis (IgA-GN) is a longterm lethal

disease that results in kidney failure if left untreated. Oxford MEST-C [349] score is an international consensus scale which can be used to evaluate IgA Nephropathy [350], evaluating different aspects: Mesangial hypercellularity (M), Endocapillary hypercellularity (E), Segmental Sclerosis (S), Tubular Atrophy (T), and Crescents (C).

The existing literature on glomeruli image modalities backed by CAD systems consider either one or two of the several aforementioned pathological features [351–353]. The pathological conclusions made under partial scores pose limits to the diagnostic and prognostic accuracy. Few works opted to investigate in-depth the possibility to build a CAD system which can be used for a full evaluation of the glomerular structures and lesions [354].

2.3.3.1 MEST-C Score

Once IgA nephropathy is diagnosed, MEST-C scoring is adopted to study the seriousness level of the disease. The MEST-C score embraces five key indicators illustrated hereunder. A reference article for Oxford classification of IgA nephropathy is available at [349].

- **Mesangioproliferation (M).** The Mesangio cells reside somewhere between the tuft and network of capillaries. The mesangioproliferation is referred to the expansion of mesangial cells characterized by pink color inside the glomerulus. The binary score is given to quantify presence of mesangial cells with 1 being more than 50% of presence in glomeruli and 0 otherwise.
- **Endocapillary Hypercellularity (E).** The Endocapillary Hypercellularity refers to the increased existence of cells in lumen of capillary which prevents the normal flow of the blood among cells by blocking the lumina. For the purpose of scoring, the binary numbers 0 and 1 indicate the absence and presence of Endocapillary Hypercellularity, respectively.
- **Glomerulosclerosis (S).** The Glomerulosclerosis affects the filters present inside the kidney which serve as blood purifiers. If these filters are blemished, the blood filtration stops which leads to kidney failure. In standard Oxford score definition, the threshold for segmental glomerulosclerosis is set to 50%, and labeled as 0 in case of tuft damaged less than half.
- **Tubular Atrophy (T).** The Tubular Atrophy is an injury that reduces the diameter of tubules by stiffening the basement membrane of the tubule. Unlike earlier scores, it is measured in terms of the occupied areas of the cortical in percentage. The values

range from 0 – 25%, 26 – 50%, and 51 – 100% of the involved areas are given score of 0, 1, and 2, respectively.

- Cellular or Fibrocellular Crescent (C). The Cellular or Fibrocellular Crescent refers to the availability of spread of proliferating capillary cells over the urinary space. The appearance of two or more extra-capillary cells in the Bowman's space denotes the crescent. The score is given 1 in case of less than 25% of damage of glomeruli and 2 in case of more than 25%, whereas 0 refers to unavailability of extra-capillary cells.

2.3.3.2 Related Works

Recently, Weis et al. [353] employed twelve Transfer learning (TL) based CNN models trained on two independent datasets and tested on third labeled dataset to distinguish nine glomerular findings. The authors were able to reach the maximum accuracy of 0.944 and the kappa value of 0.938 using ResNet-152 architecture.

Sato et al. [355] proposed a fine-tuned Neural Architecture Search Network (NAS-Net) based unsupervised framework for feature extraction and later autoencoder for evaluation using Oxford classification on images of the glomeruli of kidney biopsy.

Similarly, a CNN based InceptionV3 is modified to classify seven different pathological patterns on 15,888 glomerular images based on majority voting among nephrologists and AI system in [354]. The top performance is showed by global sclerosis with AUC for Periodic Acid-Schiff (PAS) staining equal to 0.986. However, the results of AI in segmental sclerosis, endocapillary proliferation, membrane proliferation, and crescent formation can be improved.

In [351], the detection results of Machine Learning (ML) and DL on only mesangial hypercellularity are compared considering a small dataset of 138 glomerulus images. The k -Nearest Neighbors (k -NN) and SVM are compared to AlexNet, GoogLeNet, ResNet, and Inception-v3. The modified AlexNet produced the accuracy, sensitivity, and specificity of 0.90, 0.90, and 0.80, respectively.

Another study incorporates CNN based modified model for feature extraction and SVM for the binary classification glomerulus (healthy or injured) [352]. The study was also extended to multiclass classification of particular hypercellularity lesions. The binary classification was able to reach 0.996, 0.997, 0.995, and 0.996 of precision, recall, F_1 -score, and accuracy, respectively.

Chapter 3

Radiomics and Applications of Deep Learning in Radiology

3.1 Liver Segments Classification for Surgical Planning

Liver segmentation is a pivotal task in abdominal radiology, as it is essential for several clinical tasks, including surgery, radiotherapy, and objective quantification [170]. In addition, several pieces of clinical evidence over the last few decades have demonstrated the necessity to go beyond the conception of the liver as a single object (from a morphological perspective), but rather to consider the liver as formed by functional segments [356].

Couinaud's classification [237] of liver segments is the most widely adopted system for this purpose, considering its usefulness in surgery setups. The separation of these anatomical areas can be exploited for tumor resections in CAS systems [241], allowing surgeons to focus on different segments. In this way, a surgeon can avoid removing healthy liver parenchyma, by performing a segmentectomy, also reducing the risk of complications for the patient [238].

In this section, a workflow for the segmentation of the liver segments is described. This task may be carried out directly or be preceded by the segmentation of the liver parenchyma. Although literature exists for both approaches, the experiments carried out for this thesis revealed that liver parenchyma segmentation is not needed to train a multi-class classification DL model on liver segments, since applying liver masks to the volumes reduces the context. Nonetheless, other authors found this step useful [210], therefore it may depend on the specific implementation.

3.1.1 Materials

The MSD 08 dataset [204] has been used for the purpose of training and validating the models described in this section. This dataset is composed of 443 portal venous phase CT scans. The pixel spacing in the x/y direction is in the range $[0.57, 0.9]$, whereas the thickness lies between $[2.5, 5.0]$. Liver segments' annotations ($n = 193$) have been realized in a research work by Lenovo [210]. The dataset split has been carried out as follows: 168 CT scans formed the training set and 25 CT scans composed the validation set. For the selected training set, the frequencies of class values for each segment are represented in Table 3.1. Some examples from the MSD 08 dataset with the annotations of the Lenovo team are depicted in Figure 3.1. As pre-processing, images were clipped to the Hounsfield Unit (HU) range $[-150, 350]$ and then rescaled in the range $[0, 1]$, as already performed in [56].

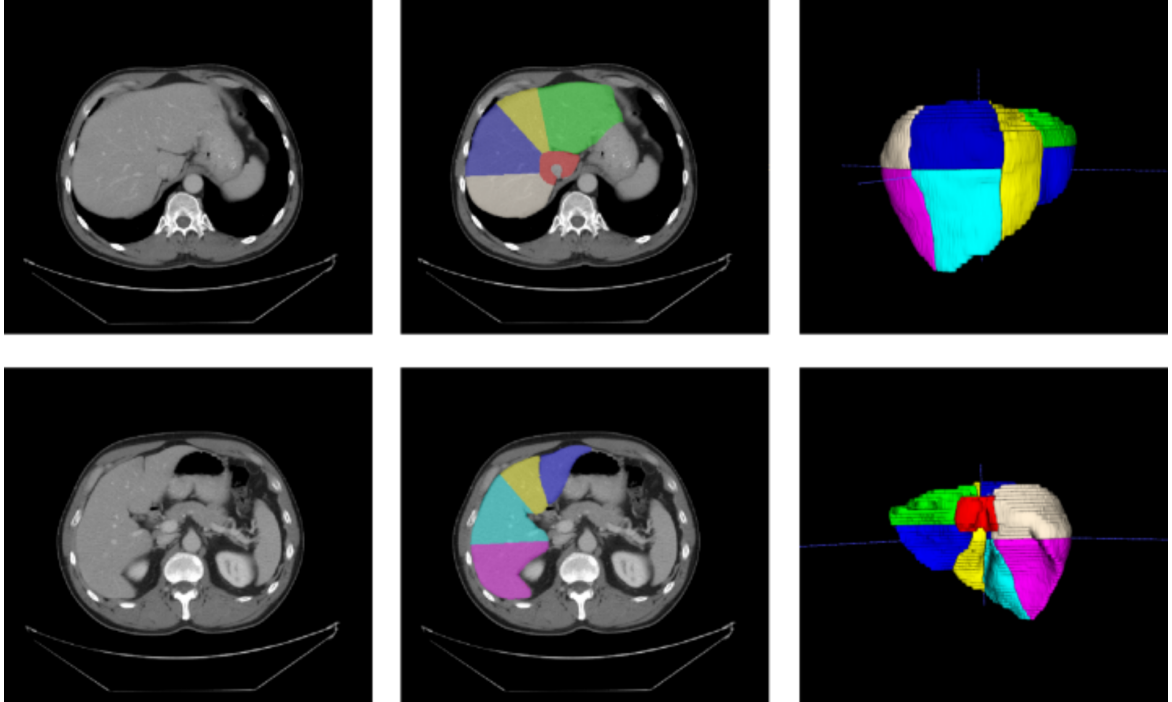


Fig. 3.1 *Medical Segmentation Decathlon Task 08 dataset*. CT scans from Simpson et al. [204], annotations from Tian et al. [210]. Left: original slices; middle: slices with segments' annotations; right: meshes' reconstructions of annotations.

Table 3.1 *Class distribution for the selected training set. Sample (n=168) from Medical Segmentation Decathlon Task 08 dataset.*

Class	Voxels	Frequency f_k (%)	$\frac{1}{f_k}$	$\log\left(\frac{1}{f_k}\right)$
Background	2,561,653,684	95.35	1.05	0.0476
Segment I	5,653,635	0.21	475.17	6.1637
Segment II	15,076,888	0.56	178.18	5.1828
Segment III	8,858,869	0.33	303.25	5.7146
Segment IV	14,139,614	0.53	189.99	5.2470
Segment V	16,275,115	0.61	165.07	5.1063
Segment VI	15,679,458	0.58	171.34	5.1436
Segment VII	23,566,210	0.88	114.00	4.7362
Segment VIII	25,548,239	0.95	105.15	4.6554

3.1.2 Workflow

The implemented segmentation architecture for the liver segments classification is a 2.5D architecture. A multi-class classification paradigm has been employed, with a focal Dice loss. The scheme of the proposed architecture is depicted in Figure 3.2.

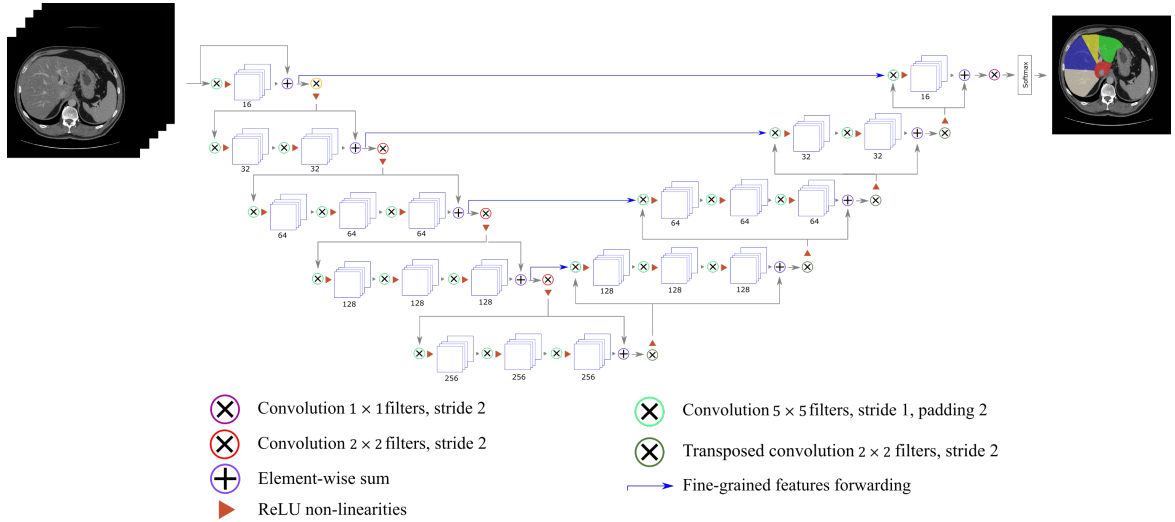


Fig. 3.2 V-Net architecture for multi-class segments' classification.

For the liver segments' classification task, 2.5D patches, i.e., 2D multi-channel, composed of 5 slices from volumes of the training set, as performed in Altini et al. [56], are randomly

sampled. The V-Net model has been trained for 1000 epochs, for each of the proposed loss function configurations reported in Section 3.1.3. The learning rate started from 0.01 and was reduced by a factor of 10 every 166 epochs. The Adam optimizer [256], with a batch size of 4, was used to carry out the training process.

3.1.2.1 Focal Loss

In order to ease the convergence process and to perform the desired optimization of the task, the loss function has to be selected accurately. Lin et al. introduced the concept of focal loss [357]. The main idea is that, for detectors (but this consideration can be extended to semantic segmentation classifiers), a mechanism is needed to focus on hard samples (i.e., samples difficult to classify), avoiding that the gradient is overwhelmed by the larger number of false negatives. According to the convention introduced in [357], the Cross-Entropy (CE) loss can be defined as:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{if } y = 0 \end{cases} \quad (3.1)$$

where $y \in \{0, 1\}$ is the binary ground truth label and $p \in [0, 1]$ is the probability estimated from the model when $y = 1$. Lin et al. introduced p_t for notational convenience:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases} \quad (3.2)$$

so that it can be written $CE(p, y) = CE(p_t) = -\log(p_t)$. A weighted version of the cross-entropy can be introduced. For instance, let α be a weighting factor in the range $[0, 1]$, so that α is adopted for class $y = 1$ and $1 - \alpha$ is exploited for class $y = 0$. Typically, α may be chosen by the inverse class frequency [357]. Introducing α_t , analogously to p_t , it is possible to define the α -balanced CE loss as:

$$CE(p_t) = -\alpha_t \log(p_t) \quad (3.3)$$

In order to redefine the cross-entropy loss, Lin et al. added a modulating factor $(1 - p_t)^\gamma$ to the CE loss, where $\gamma \geq 0$ can be found with cross-validation, and it is called the focusing parameter. Then, the CE Focal Loss (FL) is:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (3.4)$$

The original authors noted that this formulation has two main properties:

1. The loss is unaffected when there is a misclassification with a small p_t ; meanwhile, if p_t is large (near 1), the loss for well-classified examples is reduced. This means that easy examples are not penalized, if compared to “hard” examples;
2. The rate with which easy examples are down-weighted is governed by the focusing parameter γ . For $\gamma = 0$, the FL corresponds to the standard CE , whereas for larger values of γ , there is an increase in the modulating effect. Lin et al. found $\gamma = 2$ as the best value in their experiments [357].

Furthermore, FL can be weighted, so that the α -balanced FL can be defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (3.5)$$

According to [357], this yields a slight improvement in accuracy, over the naïve non- α -balanced form.

3.1.2.2 Focal Dice Loss

After the introduction of the FL , Wang et al. proposed the Focal Dice Loss (FDL) and effectively employed it for brain tumor segmentation [358].

The Dice coefficient (D_H) is usually defined for sets (let B be the binarized prediction and G the ground truth volume) as:

$$D_H = \frac{2 \cdot |G \cap B|}{|G| + |B|} \quad (3.6)$$

For the purposes of this section, D_H will be defined as the “hard” Dice coefficient, intending that it is defined for binary data only. Let us consider the estimated probability predicted volume P , composed of N voxels. Each element $p_i \in P$ ranges in $[0, 1]$, whereas each $g_i \in G$ belongs to $\{0, 1\}$. Then, the “soft” Dice coefficient D_S can be defined as:

$$D_S = \frac{2 \sum_{i=1}^N p_i g_i + \epsilon}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i + \epsilon} \quad (3.7)$$

where ϵ is a term for numerical stability, avoiding divisions by zero or by very small values. In the experiments performed, ϵ has been set to 10^{-5} . The “hard” Dice coefficient D_H is useful for assessing the final results, whereas the “soft” formulation D_S is needed during the optimization process, where a differentiable function is compulsory.

Subsequently, a “soft” Dice coefficient loss DL_S can be introduced simply as:

$$DL_S = 1 - D_S \quad (3.8)$$

These definitions hold for the binary classification case, but they can be easily generalized for $t = 1, \dots, k$ classes. For each value of t , the binary DL_{S_t} can be computed, so that the weighted multi-Dice loss (MDL_S) can be written as:

$$MDL_S = \sum_{t=1}^k \omega_t DL_{S_t} \quad (3.9)$$

where $\omega_t \geq 0$ is the weighting parameter for the class t and DL_{S_t} is the binary soft Dice loss for class t . A common way to tune the ω_t parameters is to exploit the logarithm of the inverse class frequency, as reported in Table 3.1. Nonetheless, in the experiments carried out, this did not lead to a performance enhancement, so the version with all $\omega_t = 1$, with $t = 1, \dots, k$, has been chosen. Starting from this definition, Wang et al. defined their FDL as:

$$FDL^W = \sum_{t=1}^k \omega_t (1 - D_{S_t}^{\frac{1}{\gamma}}) \quad (3.10)$$

where D_{S_t} is the “soft” Dice coefficient for class t , and a factor $\frac{1}{\gamma}$, with $\gamma \geq 1$, is applied as the exponent of D_{S_t} for each class. According to Wang et al., their FDL formulation has three properties:

1. FDL is basically unaffected by pixels, which are misclassified to class t with a large D_{S_t} . Conversely, FDL is significantly decreased when D_{S_t} is small;
2. The parameter γ regulates the rate with which better-segmented classes are weighted. For $\gamma = 1$, FDL corresponds to MDL_S . Increasing γ , the network is forced to focus on classes with bad segmentation;
3. Since FDL is based on the definition of the Dice coefficient, it focuses on the relevant metric for assessing brain tumor segmentation (the case studied in [358]).

In this thesis, two versions of the FDL have been defined. The first one is similar to the one proposed by Wang et al.:

$$FDL^A = \sum_{t=1}^k \omega_t (1 - DL_{S_t})^\gamma \quad (3.11)$$

with $\gamma \geq 1$.

The second one is based on the concept of the focal loss of Lin et al. [357]. First, the “soft” Focal Dice Coefficient (FDC) is defined. It can be formulated as:

$$FDC_{S_t} = \frac{2 \sum_{i=1}^N p_i^\gamma g_i + \varepsilon}{\sum_{i=1}^N p_i^\gamma + \sum_{i=1}^N g_i + \varepsilon} \quad (3.12)$$

Then, the FDL can be calculated as:

$$FDL^B = \sum_{t=1}^k \omega_t (1 - FDC_{S_t}) = \sum_{t=1}^k \omega_t (FDL_t) \quad (3.13)$$

The class $t = 0$ has been excluded, since it is usually associated with the background, so that the sums in Equations (3.11) and (3.13) have been divided by $k - 1$. In the proposed formulation, it is worth noting that the FDC_{S_t} numerator sums only the terms that have a $g_i = 1$. This means that two possible categories of values for p_i can be distinguished:

- A high value of p_i , associated with true positives;
- A low value of p_i , linked to false negatives.

For misclassifications, the goal of the proposed approach was to reduce the total value of FDC_{S_t} , penalizing the p_i values. Since p_i values lie in the range $[0, 1]$, a modulating factor $\gamma \geq 1$ has been added. When $p_i \leq 0.5$, it is possible to see that the reduction of p_i is greater than when $p_i \geq 0.5$. This means that a voxel with a low p_i value provides a minor contribution to the FDC_{S_t} after the modulation. It is possible to understand how the γ value affects the p_i values looking the plots in Figure 3.3.

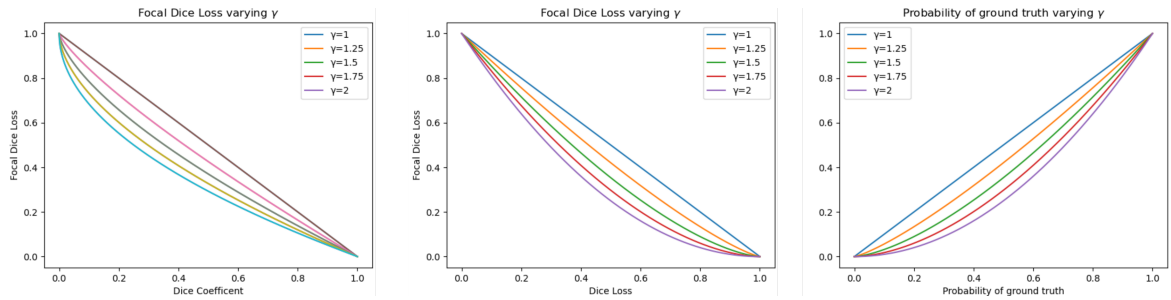


Fig. 3.3 FDL plots for the different considered formulations. Left: values of FDL^W as a function of the Dice coefficient varying γ . Center: values of FDL^A as a function of the Dice loss. Right: values of p_i varying γ in the FDL^B formulation.

3.1.2.3 Post-Processing

The post-processing consisted of two phases: in the first, only the largest connected component for each segment has been retained; in the second, all segments' inconsistencies have been eliminated, i.e., it was avoided the presence, on the same axial slice, of inconsistent segments (from an anatomical perspective). To enforce this constraint, the inconsistency matrix reported in Table 3.2 has been devised.

This table was inspired by the liver anatomy. Considering only slices on the axial plane, the liver can be subdivided into three regions: (i) above the left branch of the portal vein, (ii) below the right branch of the portal vein, and (iii) between the two branches. For each region, as it is possible to see from Figure 3.4, the allowed segments are:

1. Segments II, IV, VII, and VIII;
2. Segments III, IV, V, and VI;
3. Segments III, IV, VII, and VIII.

Table 3.2 *Segments' inconsistency table*. If the presence of the pair of segments in the same axial slice is anatomically possible, a “1” was inserted into the matrix, otherwise a “0”.

Segment	I	II	III	IV	V	VI	VII	VIII
I	1	1	1	1	1	1	1	1
II	1	1	0	1	0	0	1	1
III	1	0	1	1	1	1	1	1
IV	1	1	1	1	1	1	1	1
V	1	0	1	1	1	1	0	0
VI	1	0	1	1	1	1	0	0
VII	1	1	0	1	0	0	1	1
VIII	1	1	0	1	0	0	1	1

From these observations, the following substitution rules were derived:

- Left Branch (LB): above the LB, all the voxels classified as Segment III are relabeled as Segment II; below the LB, all the voxels classified as Segment II are relabeled as Segment III;

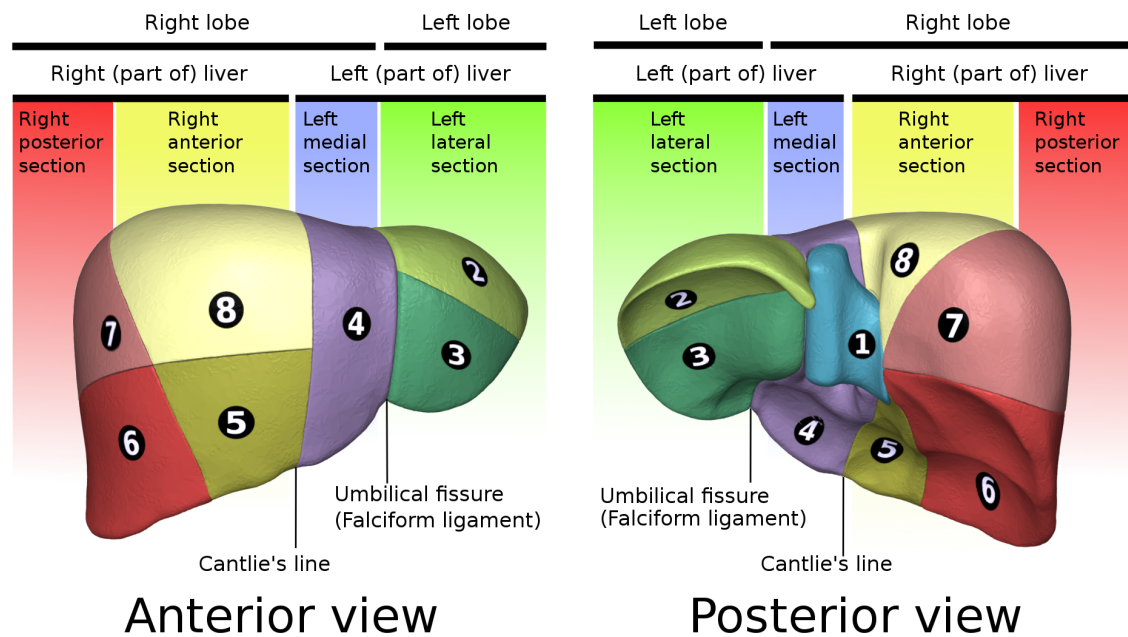


Fig. 3.4 *Liver segments' anatomy*. Data generated by the Database Center for Life Science (DBCLS) <https://dbcls.rois.ac.jp/index-en.html>, last accessed: 17 January 2022; data provided by BodyParts3D <https://lifesciencedb.jp/bp3d/?lng=en>, last accessed: 17 January 2022 [359].

- Right Branch (RB):
 - Above the RB, all the voxels classified as Segment V are relabeled as Segment VIII, and below the RB, Segment VIII labels are reassigned to Segment V;
 - Above the RB, all the voxels classified as Segment VI are relabeled as Segment VII; below the RB, all the voxels classified as Segment VII are relabeled as Segment VI.

In order to define and localize in the CT volume the two aforementioned branches, i.e., the LB and RB, of the portal vein, a majority criterion, based on the classification of the voxels, has been implemented. The first slice where the sum of the counts of the voxels belonging to Segments III, IV, V, and VI was less than the sum of the counts of the remaining voxels in the first slice above the RB.

Similarly, the first slice where the sum of the counts of the voxels from Segments II, IV, VII, and VIII was higher than the sum of the counts of the remaining voxels in the first slice above the LB.

3.1.3 Results

In order to assess the performances of the different classifiers obtained by varying the FDL formulation and the associated modulation factor γ , the “hard” Dice coefficient D_{H_t} was assessed for each class t , the average $D_{H_{avg}}$ Dice coefficient across all classes, the mean accuracy A_{avg} , and the confusion matrix, in order to see the relationships of the errors from different segments. The FDL presented by Wang et al. was compared with the ones presented in this section. Five values for the modulation parameter have been considered: $\gamma \in \{1, 1.25, 1.5, 1.75, 2\}$. The case $\gamma = 1$ is equivalent to using the Dice coefficient as defined in Equation (3.7). For each experiment, the application of the post-processing was tested to evaluate if it provides benefits or not.

Experimental results are reported in Tables 3.3, 3.4, 3.5, and 3.6, for models trained with FDL^A , FDL^B , FDL^W , and DL , respectively. Figure 3.5 shows the segmentation obtained for a single CT scan, adopting FDL^A with $\gamma = 2$.

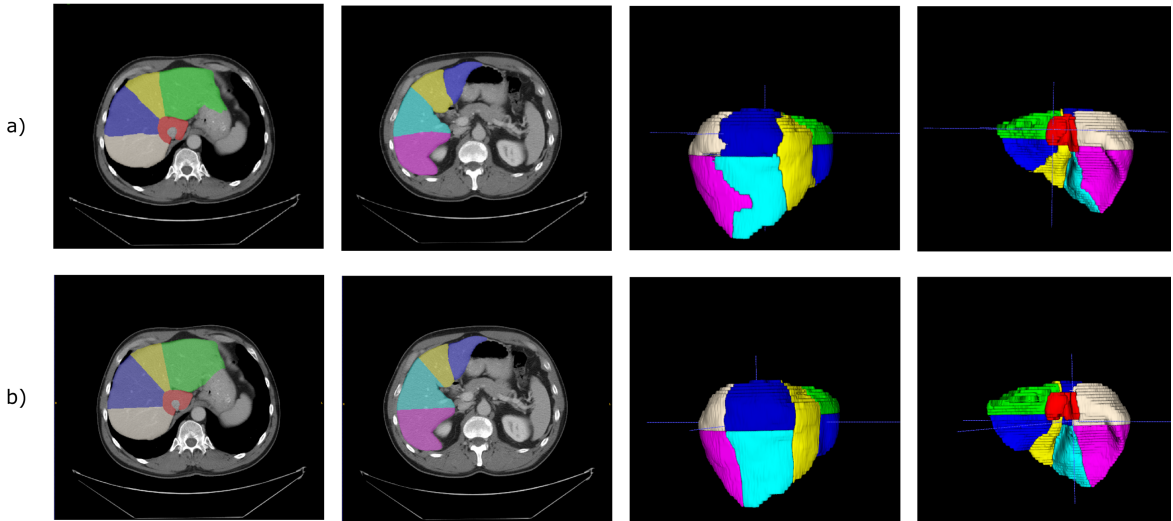


Fig. 3.5 *Qualitative results for liver segments' classification.* (a) Prediction obtained from the 2.5D V-Net trained with the proposed formulation of the focal Dice loss. For these images, the model was trained with FDL^A with $\gamma = 2$. (b) Corresponding ground truth.

In the case of FDL^W , the best model was the one with $\gamma = 1.25$ after post-processing. A possible justification is that larger values of γ result in flattening too much the loss of the volumetric predictions, so that those having a good Dice coefficient are too penalized.

For the FDL s defined in this section, the model trained FDL^A achieved the best performance also in the case $\gamma = 2$, after post-processing, while the one trained with FDL^B obtained the best performance in the configuration $\gamma = 2$.

Table 3.3 *Quantitative results for the models trained with the FDL^A formulation. FDL^A is defined as in Equation (3.11). Comparisons for the different liver segments. PP stands for Post-Processing. D_{H_t} is the “hard” Dice coefficient for class t . It is reported as a percentage. $D_{H_{avg}}$ is the arithmetic mean of D_{H_t} . $D_{H_{std}}$ is the standard deviation of D_{H_t} . The average accuracy is denoted as A_{avg} .*

Models	D_{H_0}	D_{H_1}	D_{H_2}	D_{H_3}	D_{H_4}	D_{H_5}	D_{H_6}	D_{H_7}	D_{H_8}	$D_{H_{avg}}$	$D_{H_{std}}$	A_{avg}
$\gamma = 1.25$	99.84	79.27	83.35	79.02	74.43	72.96	77.68	82.80	78.31	80.85	7.88	98.77
$\gamma = 1.25, PP$	99.84	79.27	83.35	79.02	74.43	72.96	77.68	82.80	78.31	80.85	7.88	98.77
$\gamma = 1.5$	99.84	78.82	81.04	76.75	73.88	74.62	78.50	82.19	78.92	80.51	7.74	98.75
$\gamma = 1.5, PP$	99.82	79.30	81.68	77.18	73.13	74.06	79.20	82.80	79.03	80.69	7.84	98.76
$\gamma = 1.75$	99.86	82.09	82.81	80.60	74.61	73.37	77.83	84.41	80.53	81.79	7.71	98.85
$\gamma = 1.75, PP$	99.85	81.89	83.57	80.50	74.79	73.35	78.67	84.56	80.63	81.98	7.66	98.86
$\gamma = 2$	99.88	82.31	85.38	81.56	76.84	75.58	78.91	84.96	80.76	82.91	7.17	98.92
$\gamma = 2, PP$	99.86	83.09	86.03	82.68	77.00	75.53	79.04	85.90	81.32	83.38	7.17	98.94

Table 3.4 *Quantitative results for the models trained with the FDL^B formulation. FDL^B is defined as in Equation (3.13). Comparisons for the different liver segments. PP stands for Post-Processing. D_{H_t} is the “hard” Dice coefficient for class t . It is reported as a percentage. $D_{H_{avg}}$ is the arithmetic mean of D_{H_t} . $D_{H_{std}}$ is the standard deviation of D_{H_t} . The average accuracy is denoted as A_{avg} .*

Models	D_{H_0}	D_{H_1}	D_{H_2}	D_{H_3}	D_{H_4}	D_{H_5}	D_{H_6}	D_{H_7}	D_{H_8}	$D_{H_{avg}}$	$D_{H_{std}}$	A_{avg}
$\gamma = 1.25$	99.87	82.97	83.78	80.67	72.61	74.86	79.51	83.46	79.63	81.93	7.73	98.86
$\gamma = 1.25, PP$	99.86	83.31	83.47	79.47	72.23	74.94	79.81	83.49	79.68	81.81	7.80	98.86
$\gamma = 1.5$	99.86	81.84	83.35	81.31	74.11	74.96	77.36	81.80	78.44	81.45	7.62	98.80
$\gamma = 1.5, PP$	99.85	81.89	83.93	82.21	74.28	75.54	78.45	82.48	78.77	81.93	7.47	98.82
$\gamma = 1.75$	99.85	81.89	83.93	82.21	74.28	75.54	78.45	82.48	78.77	81.93	7.47	98.82
$\gamma = 1.75, PP$	99.80	78.61	84.64	84.29	74.69	73.99	77.83	82.92	80.02	81.86	7.75	98.78
$\gamma = 2$	99.82	81.38	83.58	81.05	73.56	74.98	77.73	83.93	79.72	81.75	7.65	98.78
$\gamma = 2, PP$	99.81	80.75	83.52	80.78	73.78	75.90	78.78	84.46	80.47	82.03	7.47	98.81

Table 3.5 *Quantitative results for the models trained with the FDL^W formulation. FDL^W is defined as defined in Equation (3.10). Comparisons for the different liver segments. PP stands for Post-Processing. D_{H_t} is the “hard” Dice coefficient for class t . It is reported as a percentage. $D_{H_{avg}}$ is the arithmetic mean of D_{H_t} . $D_{H_{std}}$ is the standard deviation of D_{H_t} . The average accuracy is denoted as A_{avg} .*

Models	D_{H_0}	D_{H_1}	D_{H_2}	D_{H_3}	D_{H_4}	D_{H_5}	D_{H_6}	D_{H_7}	D_{H_8}	$D_{H_{avg}}$	$D_{H_{std}}$	A_{avg}
$\gamma = 1.25$	99.87	82.82	85.71	82.54	76.58	75.58	78.79	84.59	80.22	82.97	7.21	98.91
$\gamma = 1.25, PP$	99.87	83.03	86.18	82.89	76.73	74.83	79.01	84.61	79.81	82.99	7.33	98.91
$\gamma = 1.5$	99.85	80.49	83.84	80.10	74.14	75.43	78.04	82.75	78.65	81.48	7.57	98.79
$\gamma = 1.5, PP$	99.84	80.51	84.97	81.47	74.29	76.13	79.52	83.61	79.07	82.16	7.42	98.83
$\gamma = 1.75$	99.80	79.49	82.80	76.39	73.04	73.95	77.81	82.84	78.73	80.54	7.98	98.73
$\gamma = 1.75, PP$	99.79	79.29	82.67	76.00	73.20	74.69	78.44	83.39	79.06	80.73	7.90	98.74
$\gamma = 2$	99.79	78.92	83.57	73.08	73.35	74.76	77.80	82.45	79.74	80.38	8.18	98.73
$\gamma = 2, PP$	99.78	79.24	83.48	71.98	73.28	74.50	78.64	82.92	79.53	80.37	8.31	98.73

Table 3.6 *Quantitative results for the models trained with the Dice loss. It corresponds to the case of $\gamma = 1$ for all the previously defined focal Dice losses. Comparisons for the different liver segments. PP stands for Post-Processing. D_{H_t} is the “hard” Dice coefficient for class t . It is reported as a percentage. $D_{H_{avg}}$ is the arithmetic mean of D_{H_t} . $D_{H_{std}}$ is the standard deviation of D_{H_t} . The average accuracy is denoted as A_{avg} .*

Models	D_{H_0}	D_{H_1}	D_{H_2}	D_{H_3}	D_{H_4}	D_{H_5}	D_{H_6}	D_{H_7}	D_{H_8}	$D_{H_{avg}}$	$D_{H_{std}}$	A_{avg}
$\gamma = 1$	99.86	80.64	83.88	82.22	74.49	75.05	77.59	83.66	79.5	81.88	7.56	98.82
$\gamma = 1, PP$	99.86	80.75	84.13	83.20	74.68	75.65	77.81	83.66	79.61	82.15	7.47	98.84

Concerning FDL^B , in the experiments performed, it was found that the model trained with $\gamma > 2$ could not generalize, and the FDL_t for at least a class remained one (worst case) during the training. A possible explanation could be similar to the one provided for the FDL^W from the paper of Wang et al. [358]. When the value of γ is increased, even the high values of p_i are penalized, as it is possible to see from Figure 3.3.

A statistical test has been realized to detect if FDL^A was better than the other formulations. The Wilcoxon signed-rank test showed that FDL^A has superior performances compared to FDL^B , with $p_{value} = 0.003$, to FDL^W with $p_{value} = 0.034$, and to DL with $p_{value} = 0.011$. Considering, as usual, the significance threshold $\alpha_{th} = 0.05$, the obtained results were statistically significant.

The normalized confusion matrices are reported in Figure 3.6. Elements in the main diagonal of the row-normalized and column-normalized confusion matrices are the *precision(s)* and *recall(s)* of the corresponding classes, respectively. From the confusion matrices, it is possible to see that the most frequent errors arose between (in all the list here, the notation is predicted segment and ground truth segment): Segment I and the background (this could be due to the fact that the CNN classifies the vena cava as Segment I, instead of background), Segment II and Segment III (this depends on the fact that the LB is predicted in a lower position than the real one), Segment IV and Segment II (they are adjacent on the axial plane), Segment VIII and Segment IV (they are adjacent on the axial plane), Segment V and Segment VI (they are adjacent on the axial plane), and Segment VIII and Segment VII (they are adjacent on the axial plane).

Existing approaches are compared in Table 3.7. It is worth noting that existing works did not provide results (as D_H) for each segment. Different research works, such as [239, 241], only proposed a qualitative assessment of the obtained results via expert radiologists, whereas Tian et al. [210] and Yan et al. [257] only reported the aggregated Dice coefficient per case across all classes.

3.1.4 Summary

In Section 3.1, two novel focal Dice loss formulations are proposed for liver segments' classification and compared with the existing literature. The problem of correctly identifying liver segments is pivotal in surgical planning since it can allow more precise procedures, such as executing segmentectomies. In fact, according to the Couinaud model [237], hepatic vessels are the anatomic boundaries of the liver segments. Therefore, in current hepatic surgery, the identification of the liver vessel tree can avoid the unnecessary removal of

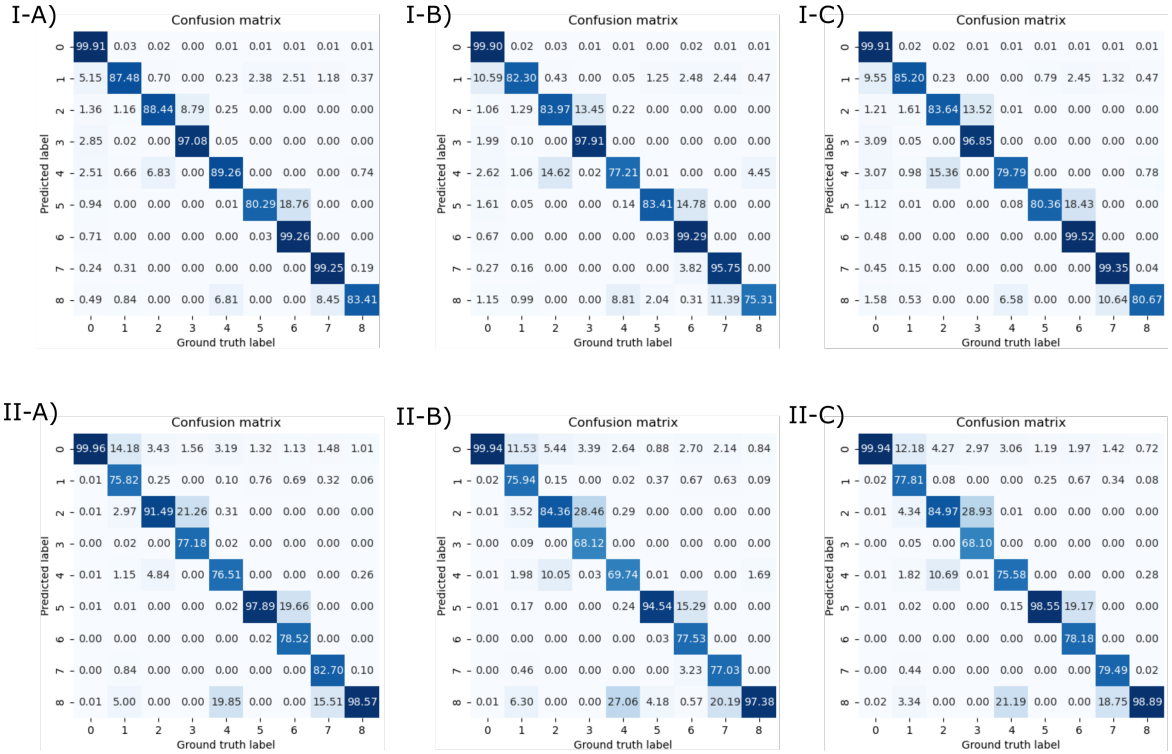


Fig. 3.6 *Normalized confusion matrices*. **(Top)** Row-normalized confusion matrices. **(Bottom)** Column-normalized confusion matrices. Left: Confusion matrix for the model trained with FDL^W . Center: Confusion matrix for the model trained with FDL^A . Right: Confusion matrix for the model trained with FDL^B .

Table 3.7 *Literature overview for liver segments' classification.* Quantitative results are not available for [239, 241], since they only provided a visual assessment of the segments' classification.

Paper	Method	Dataset	Dice per Case
Tian et al. [210]	GLC U-Net	MSD 08, n=193	92.80 ± 3.80
	U-Net		83.86 ± 5.44
	U-Net 2.5D		89.04 ± 4.30
	U-Net 3D		88.93 ± 4.44
	U-Net with convolutional LSTM		90.06 ± 4.01
Yan et al. [257]	FCN [75]	Local Dataset, n=500	82.60 ± 4.29
	U-Net [72]		83.80 ± 4.01
	U-Net++ [85]		82.69 ± 4.57
	nnU-Net [258]		81.70 ± 4.12
	AU-Net [259]		82.31 ± 4.23
	PraNet [260]		83.03 ± 4.28
Oliveira et al. [239]	Geometric fit algorithm	SLIVER07 [126], n=20	N/A
Yang et al. [241]	Nearest Neighbor Algorithm	Local Dataset, n=43	N/A

healthy liver parenchyma, thus lowering the risk of complications that can arise from larger resections [238]. Despite this, in the literature, many authors that focused on liver segments' classification methods usually only provided visual assessment (as [239, 241]), making it difficult to benchmark different algorithms for this task. Thanks to the work of Tian et al. [210], a huge dataset composed of CT scans with liver segments' annotation is now publicly available, easing the development and evaluation of algorithms suited for the task. The proposed CNN-based approach allowed the achievement of reliable results, with the best formulation, FDL^A with $\gamma = 2$, obtaining an average Dice coefficient higher than 83% and an average accuracy higher than 98%. The adoption of rule-based post-processing, which implements a mechanism for detecting and resolving inconsistencies across axial slice images, allowed the achievement of a slight enhancement of the obtained results for different models. Given that the difference between the models was exiguous, it may be necessary to exploit a different architecture to assess a clearer difference between these loss functions.

A limitation of this work is the absence of an external validation dataset, which is due to the fact that there is a lack of annotated datasets with liver segments and it is very expensive to create a local one labeled by radiologists. Furthermore, with greater GPU memory disposable,

more complex 3D architectures capable of handling a larger context can be exploited. Instead, in this work, the context was limited to five slices at once.

As a possible future direction, it is worth noting that another possible way to approach this task consists of determining regression planes using a 3D CNN. Even if there are classical approaches (as Oliveira et al. [239]) devoted to the estimation of these cutting planes, there are actually no DL-based approaches in this direction (Tian et al. [210] and Yan et al. [257], developed a semantic segmentation model too), which can be more resilient to overfitting and to the prediction of inconsistent segments' classification with respect to voxelwise predictive models, as those based on semantic segmentation.

3.2 Vertebrae Segmentation and Identification

The work described in this section aims to propose a novel algorithm for vertebrae identification, which is simpler than more sophisticated methods already proposed in the literature. In addition, the advantage of the proposed approach is that it does not require single vertebrae-level annotations to be trained. A method for binary spine segmentation based on 3D FCNs is also developed in this section. Finally, a visualization tool has been implemented to assess the results of the considered methodologies qualitatively.

The method developed in this section fuses traditional machine learning techniques with DL ones to achieve state-of-the-art results. The availability of a great deal of spine CT datasets enables the possibility to train DL models for the spine segmentation task. The proposed two-fold approach first exploits a 3D CNN that automatically segments the whole spine; subsequently, traditional machine learning algorithms take the responsibility to locate centroids in the final stage. The k -means algorithm with morphological operators and shape descriptors analysis, which starts from the binary segmentation results, enables recovering the masks of the individual vertebrae. This method permits achieving respectable results without needing any single vertebrae-level annotations for training.

One of the limitations of the proposed methodology is its semi-automaticity. However, it offers potential benefits over simple 3D component analysis of the segmented region, as it may result in the spine being considered a unique connected component. In addition, the unavailability of a dataset large enough is always a considerable problem; thus, the proposed methodology can be useful also when there is little or no annotated data. A tabular comparison of the proposed approach with existing research methods is listed in Table 3.8. Further comparison information can be found in later sections.

Table 3.8 *Comparison between the proposed method and related works*. Results are reported in terms of Dice coefficient (*DSC*). CT stands for computed tomography, CNN stands for convolutional neural network, PaDBN stands for patch-based deep belief networks, PCNN stands for pulse coupled neural network, APCNN stands for adaptive pulse coupled neural network, MLPNN stands for multi-layer perceptron neural network and MLPNN1f means MLPNN considering only the intensity level of each voxel as a feature.

Reference	Method	Test Sample	<i>DSC</i> [%]
Proposed	3D V-Net	50 CT scans	89.17 ± 3.63
Kim et al. [270]	U-Net	14 CT scans	90.40
Vania et al. [263]	CNN	32 CT scans	94.28 ± 3.25
Qadri et al. [271]	PaDBN	3 CT scans	86.1
Lessmann et al. [269]	3D U-Net	25 CT scans	84.6 ± 6.9
Zareie et al. [273]	PCNN	17 CT scans	65.7 ± 15.4
	APCNN	17 CT scans	95.0 ± 2.3
	MLPNN	17 CT scans	91.1 ± 2.9
	MLPNN1F	17 CT scans	77.3 ± 4.7
	APCNN (noise 3%)	17 CT scans	94.3 ± 2.6
	MLPNN (noise 3%)	17 CT scans	87.8 ± 4.1

3.2.1 Materials

The main publicly available datasets concerning spine segmentation are: VerSe‘19 and VerSe‘20 [261, 266, 267], CSI-Label 2014 [360, 361], CSI-Seg 2014 [265, 268] and xVertSeg [264]. The latter three are listed by SpineWeb (<http://spineweb.digitalimaginggroup.ca/>, *last accessed: 6 June 2021*), an important online archive for multi-modal spine imaging data. A summary of the available datasets is presented in Table 3.9. xVertSeg is a collection of 25 lumbar-only CT scans with voxel-level annotations that include fractured vertebrae. CSI-Seg 2014 and CSI-Label 2014 have become available during the MICCAI 2014. While the former includes segmentation masks, the latter is provided with centroid annotations.

The VerSe‘19 and VerSe‘20 challenges enabled the adoption and benchmarking of deep learning-based techniques for spine segmentation since they have an adequate sample size with a variety of conditions, fields of view, and labeled vertebrae. For instance, the VerSe‘19 data include a variety of fields of view (e.g., cervico-thoraco-lumbar and thoraco-lumbar scans), a compound of isotropic and sagittal reformations, and subjects with metallic implants or vertebral fractures. The VerSe‘20 dataset includes atypical anatomies such as transitional vertebrae and other vertebrae, i.e., L6, sacralization of L5 and C7 with cervical ribs. For running the experiments, a dataset of 214 spine multi-detector CT scans

has been extracted from VerSe‘20 challenges data. The images can be downloaded from the OSF repository (<https://osf.io/t98fz/>, *last accessed: 6 June 2021*) and are available in Neuroimaging Informatics Technology Initiative (NIfTI) format. The dataset is split as follows: 148 CT scans for training, 16 CT scans for validation, and 50 CT scans as the final test set. 12 CT scans collected from Medica Sud s.r.l. were also considered, in order to assess the performance of the vertebrae labeling algorithm on patients with scoliosis, ranging from mild to severe cases.

Table 3.9 *Spine segmentation datasets*. In the labels column, C stands for centroids’ labels, M for masks, and S for scoliosis severity. Medica Sud s.r.l. (<https://www.medicasud.it/>, *last accessed: 6 June 2021*) is a local medical clinic that provided 12 CT scans. Please note that this dataset is not publicly available.

Dataset	Spine Tract	Sample Size	Modality	Labels
xVertSeg [264]	Lumbar	$n = 25$	CT scans	C
CSI-Seg 2014 [265, 268]	Thoraco-lumbar	$n = 20$	CT scans	M
CSI-Label 2014 [360, 361]	Whole spine	$n = 302$	CT scans	C
Verse‘19 [261, 266, 267]	Whole spine	$n = 160$	CT scans	C + M
Verse‘20 [261, 266, 267]	Whole spine	$n = 300$	CT scans	C + M
Medica Sud s.r.l.	Whole spine	$n = 12$	CT scans	S

3.2.2 Workflow

To ensure the reproducibility of the algorithms introduced in Sections 3.2.2.1 and 3.2.2.2, and the visualization tool presented in Section 3.2.2.3, the code has been made publicly available on GitHub (https://github.com/Nicolik/Segm_Ident_Vertebrae_CNN_kmeans_knn, *last accessed: 6 June 2021*).

3.2.2.1 Spine Segmentation

The images have been pre-processed according to the method proposed by Payer et al. [276], which consists of reorienting, smoothing, and clamping. In the work herein described, the clamping has been performed in the range $[-150, 1000]$, instead of $[-1024, 8192]$, since high atomic number structures such as bone have HU values in the range $[250, 1000]$ [362]. Images and the related masks have been cut with the smallest bounding box containing the spine. Images have been resampled to an isotropic resolution of 1 mm, as in the work from Payer et al. [276]. The workflow followed for the spine segmentation is depicted in Figure 3.7.

The binary segmentation stage is performed by exploiting the V-Net architecture proposed by Milletari et al. [71]. The Dice loss function formulation adopted is the same adopted in the work of Altini et al. [56]. V-Net is an encoder–decoder architecture. The first part of the network is devoted to the feature extraction process; the second reconstructs high resolutions masks, exploiting also skip connections across the encoder–decoder paths. Compared to U-Net, it is worth noting that V-Net exploits down-convolutions, with stride $2 \times 2 \times 2$ and kernel-size $2 \times 2 \times 2$, instead of fixed downsampling realized by $2 \times 2 \times 2$ max-pooling, adding trainable parameters also in this stage. As differences between the original V-Net architecture and that implemented for this thesis, there is the adoption of ReLU non-linearities instead of PReLU ones [45] and the adding of a Batch-Normalization (BN) layer after each convolutional layer, as also done in Altini et al. [56] and Shen et al. [86].

The network has been trained for 500 epochs with data augmentation (random noise with $\mu = 0, \sigma^2 \in [0, 0.1], p = 0.1$). Every 100 epochs, the trained model has been saved to check for overfitting issues. With the aim to implement the patch-based pipeline and the augmentation operation, the TorchIO library has been utilized [224]. The parameters for the patch sampler were: uniform sampling, 8 patches per volume, and patch size of $64 \times 64 \times 64$. The original V-Net architecture processed input volumes of $128 \times 128 \times 64$, but it was decided to limit the dimension for allowing larger batches to benefit from BN layers.

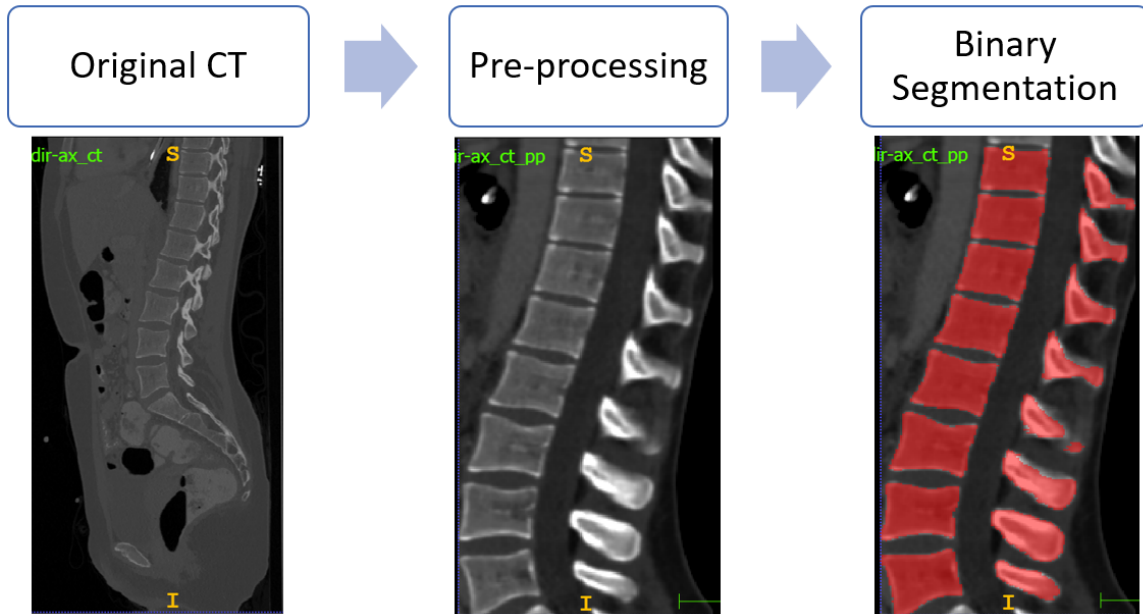


Fig. 3.7 *Vertebrae segmentation workflow.*

3.2.2.2 Vertebrae Identification

Determining vertebrae centroids is a prerequisite for achieving vertebrae identification. Indeed, centroids can be then exploited to recover single vertebrae masks. This task has to be carried out after the binary segmentation of the spine. A 3D connected component analysis of the segmented region will not solve the problem of vertebrae labeling, since the spine will likely form a unique connected component. For this reason, the chosen approach is a semi-automated workflow that consists of different steps, two of which require input from the user. The workflow followed for vertebrae identification is depicted in Figure 3.8.

The sequence of the involved operations is listed below:

- **Vertebrae Number Selection.** This step requires input by the user, which has to insert the number of visible vertebrae given the binary segmentation. The user has to provide the name of the first vertebra (from the top to the bottom) in order to perform the correct labeling according to the legends provided by VerSe.
- **Slice Extraction.** The algorithm extracts a 2D sagittal slice from the binary segmentation, starting from the middle of the image since it has a higher probability of showing well-clustered vertebrae. Kindly note that, in some cases, e.g., patients affected by severe scoliosis, this consideration may not hold, resulting in lower segmentation performances. Inside the selected slice, the following sub-steps have been carried out:
 - **Morphological and Connected Components Analysis.** The morphological analysis aims at removing small points which can be wrongly considered as standalone components, whereas the purpose of the connected components' analysis is to label each component with a different value.
 - **Shape Descriptor and Clustering for Arches and Bodies.** It is worth noting that every single component is either a vertebral body or a vertebral arch, so it is important to correctly assign each component to the appropriate category. This stage carries out the above process by considering the proper shape descriptors of the individual components.
 - **Arch/Body Coupling.** This step connects each vertebral arch to the nearest vertebral body, by assigning the same label to both.
 - **Centroids' Computation and Slice Showing.** If the output vertebrae number matches the input number from the first step, the algorithm goes further with the computation of centroids' positions for each vertebra; otherwise, the process has to be repeated from another slice.

- **Best Slice Selection and Centroids' Storage.** The algorithm repeats the workflow until it reaches a slice without connected components. Then, the user chooses the best slice among the displayed ones, and the algorithm stores the centroids' position.
- **3D Multi-class Segmentation.** Centroids are used in a k -NN classifier to produce a 3D segmentation map in which each vertebra has its own label.

In the next paragraphs, the non-trivial steps of this process are described.

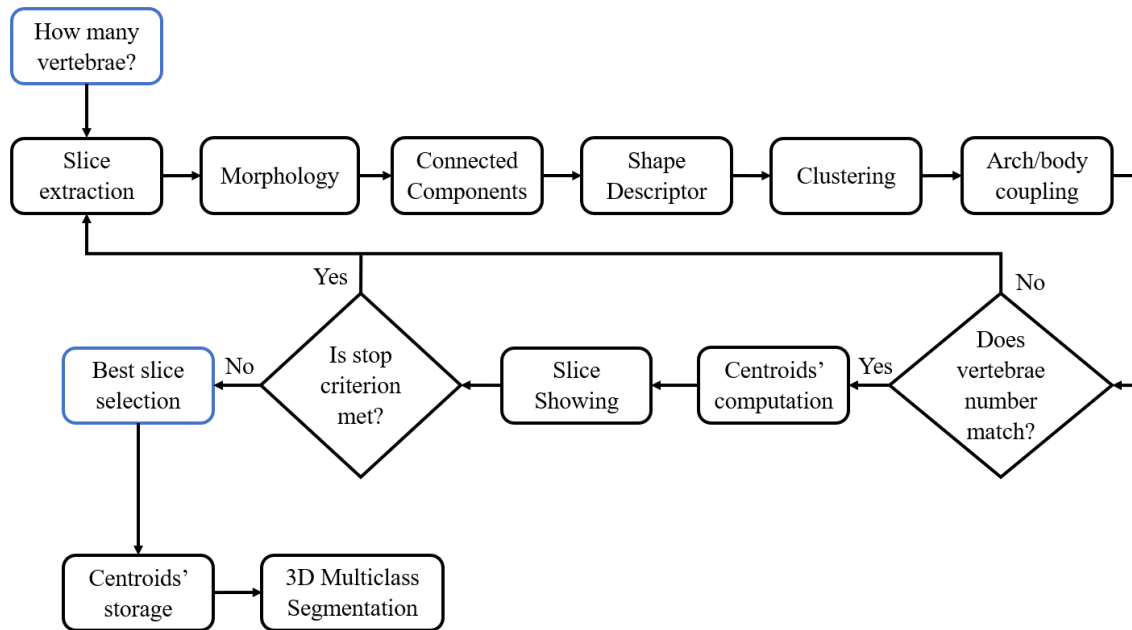


Fig. 3.8 *Vertebrae identification workflow*. Steps that require user interaction have blue borders.

Morphological and Connected Components Analysis As mentioned in Section 3.2.2.2, the morphological analysis is targeted at removing outliers, i.e., small points that are not connected to vertebral arches or bodies. To carry out this process, the *ErodeObjectMorphologyImageFilter* and the *DilateObjectMorphologyImageFilter* were exploited, which are provided by the SimpleITK library.

The former filter erases every component contour, whereas the latter enlarges the component contours. It means that if a component is relatively small, erosion will completely remove it; the combination of both removes the small components without affecting the bigger ones. The morphological analysis has been performed with a kernel radius of 2 for both erosion and dilatation.

Figure 3.9 shows the effects of the morphological analysis.

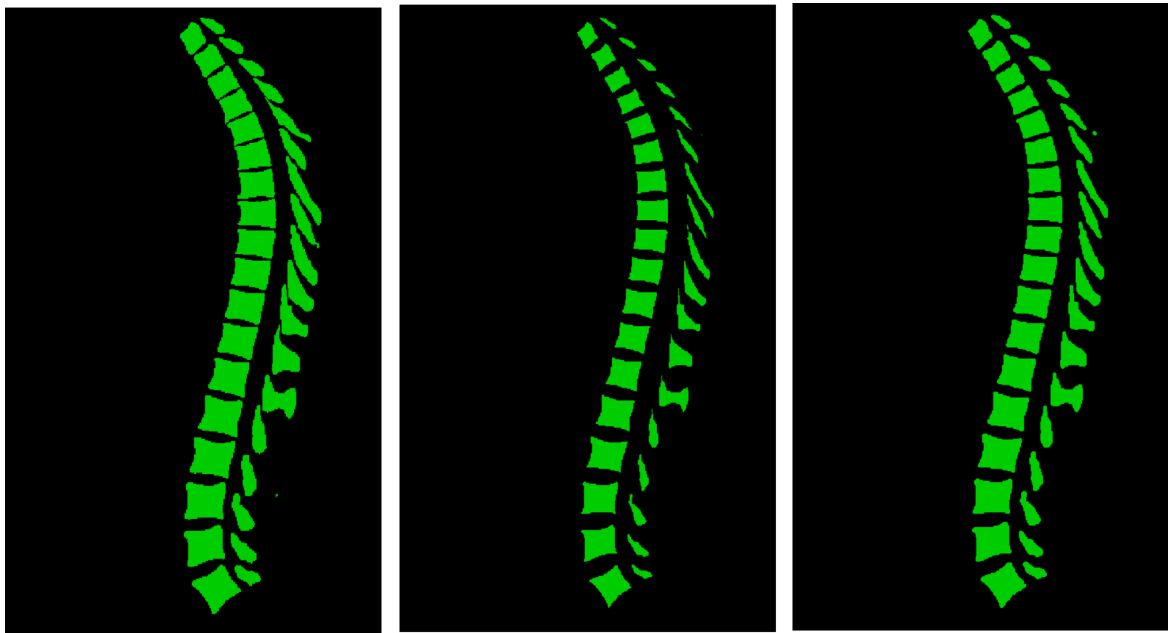


Fig. 3.9 *Workflow of the morphological analysis*. Note that small components are removed in the erosion phase, and they do not appear in the final output.

The connected components analysis substage employs the SimpleITK library too, which offers the *ConnectedComponentImageFilter* that labels each component with a different intensity value. Figure 3.10 shows the output of this filter.

Shape Descriptor and Clustering This step creates a shape descriptor for each component so that arches and bodies can be clustered in different subsets.

The shape descriptor sub-stage exploits the Scikit-image library that provides the *regionprops* and the *regionprops_table* methods which automatically compute several descriptors for each component. Some of these descriptors are used to set up a Pandas (<https://pandas.pydata.org/>, last accessed: 6 June 2021) *DataFrame* which is used to group all similar components in two clusters, i.e., arches and bodies. The chosen feature set is composed of the following descriptors (definitions in italics are taken from the Scikit-image documentation):

- **Area:** *The number of pixels of the region.*
- **Centroid:** *The centroid's position of the region.*
- **Extent:** *Ratio of pixels in the region to pixels in the total bounding box of the region.*

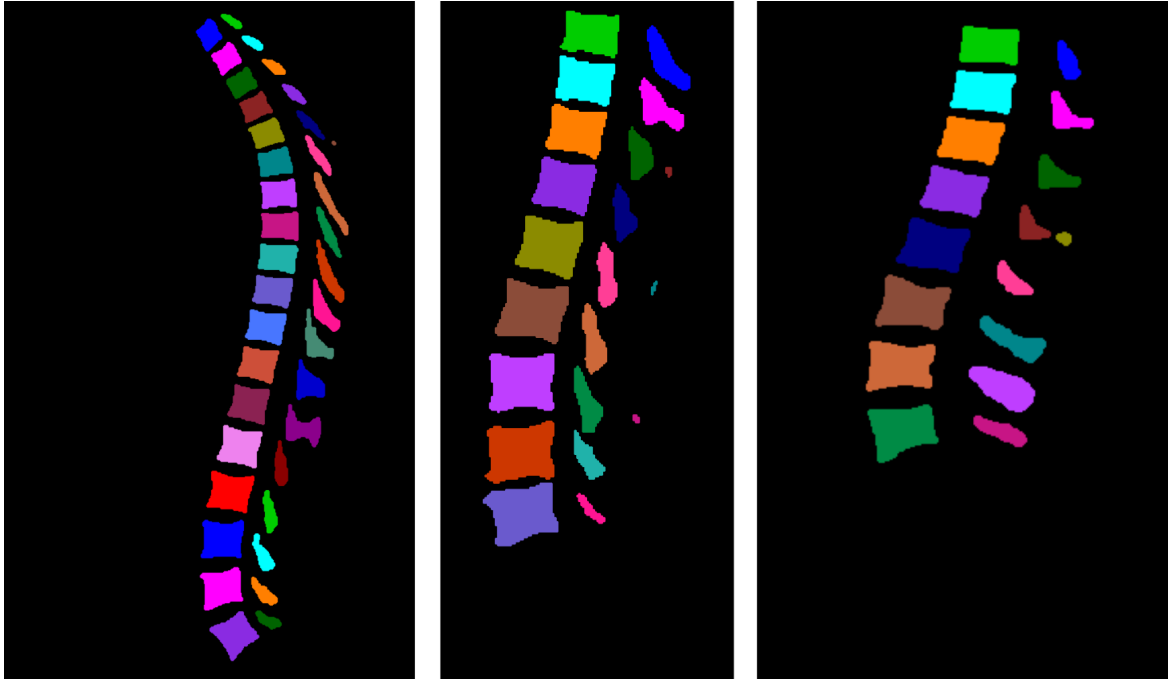


Fig. 3.10 Sample *ConnectedComponentImageFilter* output.

- **Perimeter:** *Approximation of the perimeter of the region.*
- **Eccentricity:** *Ratio of the focal distance (distance between focal points) over the major axis length.*
- **Solidity:** *Ratio of pixels in the region to pixels of the convex hull image (the smallest convex polygon that encloses the region).*

These features have been selected with a heuristic process, choosing the subset which better discriminates bodies and arches. It is also worthy of note that, for scoliosis patients, adding **Inertia Tensor** (*Inertia tensor of the region for the rotation around its mass*) to this feature set leads to an improvement in the results.

The *k*-means clustering algorithm was exploited for realizing the clustering of vertebral arches and bodies. The drawback is that it is sensible to outliers; however, after the morphological analysis, the probability of retaining outliers is largely reduced. The *k*-means clustering is a popular algorithm for cluster analysis that finds commonalities in data and groups them without the need for ground truth. Cluster analysis, in fact, belongs to the unsupervised learning branch of machine learning. At the start of the algorithm, *k* centroids are initialized (randomly, from *k* data points or from other prior knowledge) in the feature hyperspace. Then, every sample is assigned to the nearest centroid according to some distance

metric. Centroids for the next iteration are computed as the average of the coordinates of the points of each cluster. The process is repeated until convergence [363].

The Scikit-learn library provides its own implementation of the k -means clustering model. The only parameter to tune is the number of clusters, which is set to two. Before fitting the model, it is important to normalize the *DataFrame* using the Z-score normalization. The main hyper-parameter to tune for k -means clustering is the number of clusters. Since the purpose is to distinguish between arches and bodies, the *n_clusters* hyper-parameter has been set to 2. Other hyper-parameters are set to their default values. Experiments where the *init*, *n_init*, *algorithm* and *max_iter* hyper-parameters were changed did not result in improvements.

After performing the clustering, it is essential to determine which cluster regards vertebral bodies and which cluster concerns vertebral arches. This step is required because the cluster analysis does not know the nature of the samples, but only produces the two groups. For this purpose, the distinction is made by looking at the total area of the clusters: the greatest one represents the vertebral bodies. The cluster analysis is not always successful: cervical vertebrae's shape is not well distinguishable from the arch, and this can lead to an inaccurate outcome.

Arch/Body Coupling As mentioned before, this step assigns each vertebral arch to the nearest vertebral body. For each vertebral arch, the Euclidean distance between the arch and every vertebral body is computed. Let N_v be the number of vertebrae and N_a the number of segmented arches. Then, the sets of vertebral arches and bodies can be denoted as $\{a_j\}_{N_a}$ and $\{v_i\}_{N_v}$, respectively, with $j = 1, \dots, N_a$ and $i = 1, \dots, N_v$. Distance between pairs of a_j, v_i can be defined as in Equation (3.14).

$$d(a_j, v_i) = \sqrt{(a_{jx} - v_{ix})^2 + (a_{jy} - v_{iy})^2 + (a_{jz} - v_{iz})^2} \quad (3.14)$$

First, the distance between every a_j and v_i is computed. Then, iteratively, the vertebral body which has the minimum distance from the arch is linked to the vertebra body, and the labels are merged. Figure 3.11 depicts the output of cluster analysis and coupling.

Multi-class Segmentation Once the centroids have been computed, every point of the volume which is not part of the background is assigned to the same label as the nearest centroid according to Euclidean distance. To accomplish this procedure, a non-parametric method, the k -NN classifier, which is based on distances between samples, was exploited. The workflow of this algorithm can be summarized into three main steps [364]:

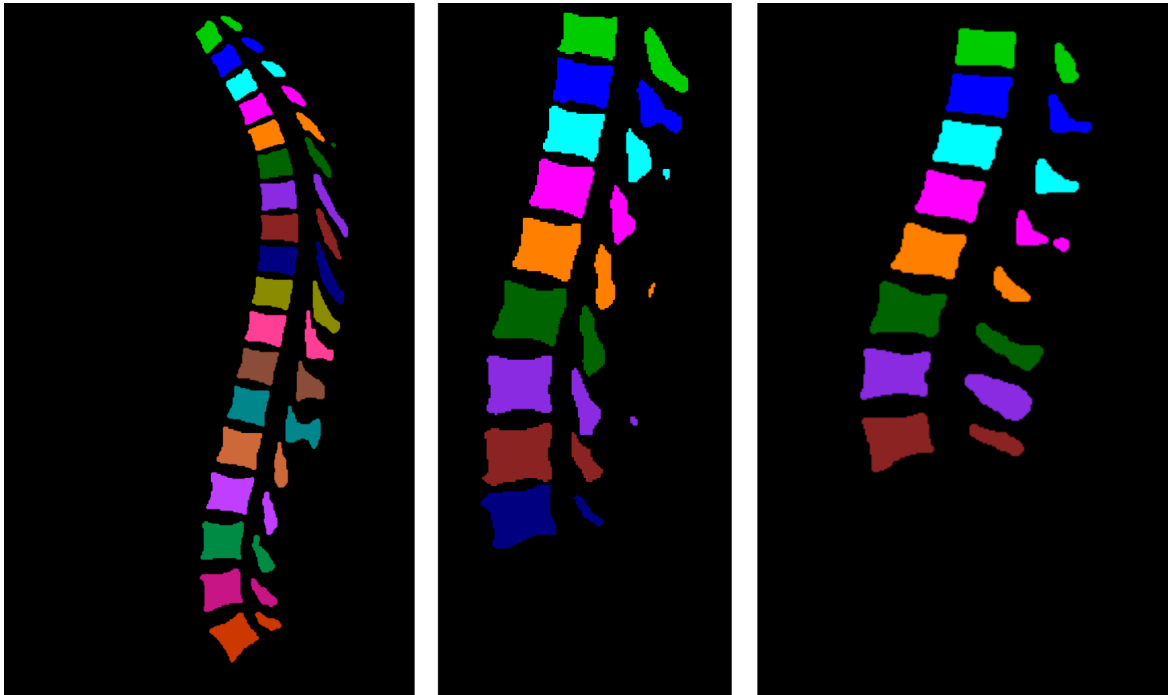


Fig. 3.11 *Sample outputs of cluster analysis and arch/body coupling.*

- The learning phase, which is not mandatory, results in the partitioning of the hyperspace in clusters based on samples' positions.
- The distance computation phase consists of the computation of all the distances between samples and centroids (the most used distance metric is the Euclidean Distance, as in the proposed pipeline, but it is also possible to use Manhattan Distance or other distances).
- The classification phase assigns each sample to the class of the nearest cluster's centroid.

For the purposes of this research, the learning phase has not been performed since the centroids have already been computed in the k -means clustering stage.

3.2.2.3 Visualization Tool

In this study, a visualization tool, that enables displaying the CT scans and masks through a graphic user interface (GUI) based on the Qt framework, ITK, VTK, and OpenCV libraries [365–367], was also developed. It offers multiple functionalities, such as image contrast adjustment, mask visualization, and mesh reconstruction.

As depicted in Figure 3.12, the interface is composed of three major parts. The toolbox section facilitates the loading and reading of the CT images from a local folder and permits smoothly navigating through the slices. The toolbox comprises a windowing option to set custom values of window-width and window-level and adjusts image contrast for optimized visualization of the anatomical regions of interest. The segmentation results are overlaid on the original images and each vertebra is colored differently so that identifying vertebrae is convenient for the user. Furthermore, the vertebrae's centroids and names are shown in the sagittal and coronal views. Three of the four windows in the views section show the CT scan in the axial, sagittal, and coronal planes. A fourth window is used to show a reconstruction of the volume obtained from the segmentation masks. The aforementioned task is achieved by using the VTK discrete marching cubes algorithm.

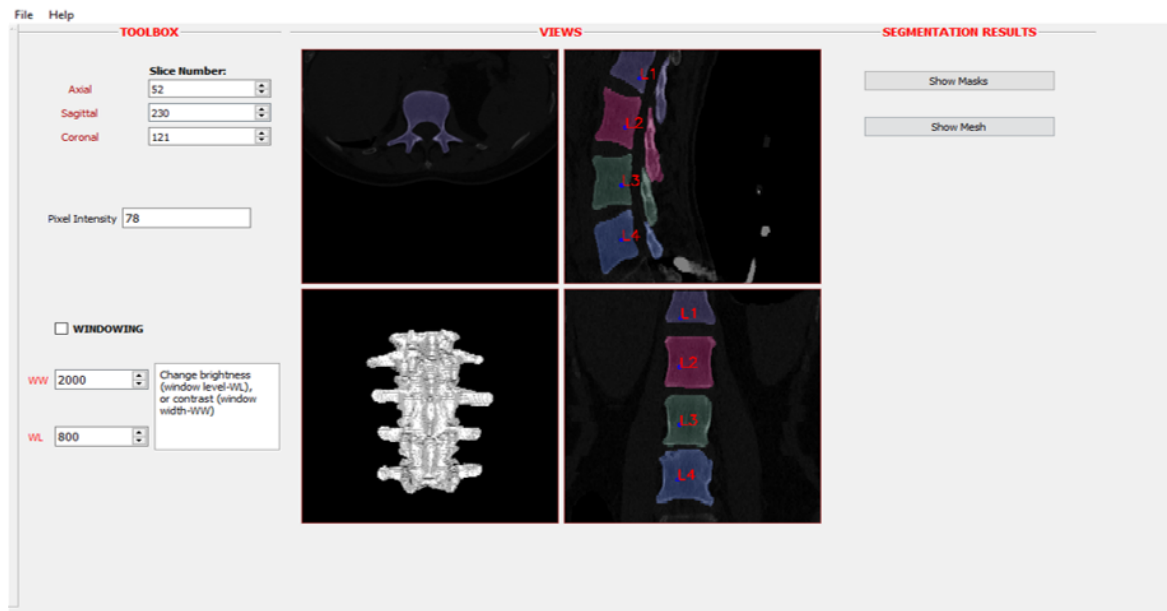


Fig. 3.12 Visualization tool for visualizing CT scans and masks.

3.2.3 Results

Regarding the binary segmentation task, 3D V-Net obtained the results reported in Table 3.10 considering a test set of 50 images. Quality measures are reported for the CNN trained every 100 epochs, showing that convergence for the considered model can be achieved after 200 epochs. These results allowed the achievement of a realistic segmentation of the whole spine, as can be seen from Figure 3.13.

Table 3.10 *Spine binary segmentation results*. *ASSD* stands for average symmetric surface distance, whereas *MSSD* stands for maximum symmetric surface distance. The best result for each of the reported measures is in bold font.

Epochs	DSC [%]	Recall [%]	Precision [%]	ASSD [mm]	MSSD [mm]
100	85.07 $\pm$ 3.02	94.25 $\pm$ 3.79	77.65 $\pm$ 3.78	2.46 $\pm$ 0.81	63.16 $\pm$ 23.27
200	88.20 $\pm$ 2.66	93.46 $\pm$ 4.03	83.61 $\pm$ 2.65	1.89 $\pm$ 0.61	60.87 $\pm$ 23.46
300	88.44 $\pm$ 2.69	93.85 $\pm$ 4.49	83.78 $\pm$ 2.81	1.91 $\pm$ 0.56	64.03 $\pm$ 28.96
400	88.34 $\pm$ 2.35	94.51 $\pm$ 3.31	83.02 $\pm$ 2.81	1.85 $\pm$ 0.63	62.77 $\pm$ 27.67
500	89.17 $\pm$ 3.63	93.60 $\pm$ 6.27	85.43 $\pm$ 2.75	1.43 $\pm$ 0.63	56.69 $\pm$ 18.07

For what concerns the multi-class vertebrae identification, a multi-class *DSC* of $90.09 \pm 3.14\%$ was obtained. The multi-class *DSC* score for the vertebrae labeling task was calculated in an ideal condition, i.e., starting the vertebrae identification procedure from ground truths of spine segmentation. Results are reported as mean $\pm$ standard deviation. Multi-class *DSC* is computed as the average of the binary *DSC*s for each vertebra class. Kindly note that the vertebrae identification algorithm proposed in this section may not work optimally for the C1, C2, and C3 vertebrae, since the shapes of the bodies of these vertebrae can be easily mistaken with their arches. The centroids' predictions examples are shown in Figure 3.14. The following vertebrae labeling, obtained with *k*-NN, is depicted in Figure 3.15.

In order to assess the quality of the vertebrae labeling stage on scoliosis cases, a collection of 12 CT scans coming from Medica Sud s.r.l. was exploited. Four out of the 12 CT scans belong to patients with severe scoliosis, 4 to patients with moderate scoliosis, and the remaining 4 to patients with mild scoliosis. For these images only information about the severity of the scoliosis was available. Therefore, the method of Payer et al. [276] was exploited as the gold standard for the segmentation masks. The multi-class *DSC* has been considered as a similarity measure across the two methodologies. The same binary spine segmentation was achieved for both algorithms, obtained by exploiting Payer's pre-trained models, therefore focusing the comparison on the labeling stage. Multi-class *DSC*s have been determined separately for each scoliosis category, obtaining $43.40 \pm 30.03\%$, $70.61 \pm 18.50\%$, and $83.38 \pm 12.51\%$, for severe, moderate, and mild scoliosis patients, respectively. After adding **Inertia Tensor** to the feature set for discriminating between arches and bodies, multi-class *DSC*s grown, managing to obtain $49.79 \pm 24.90\%$, $77.76 \pm 15.05\%$, and $83.48 \pm 12.56\%$. These results show a clear improvement for severe and moderate scoliosis cases.

The proposed vertebrae identification stage appears to have promising results on mild and moderate scoliosis; however, the identification phase for severe scoliosis cases can be further improved in future works. This problem arises from the fact that the *k*-NN model has

a too low complexity for modeling severe scoliosis cases. Example results on this dataset are shown in Figure 3.16.

A comparison between the proposed method and related works is reported in Table 3.8.

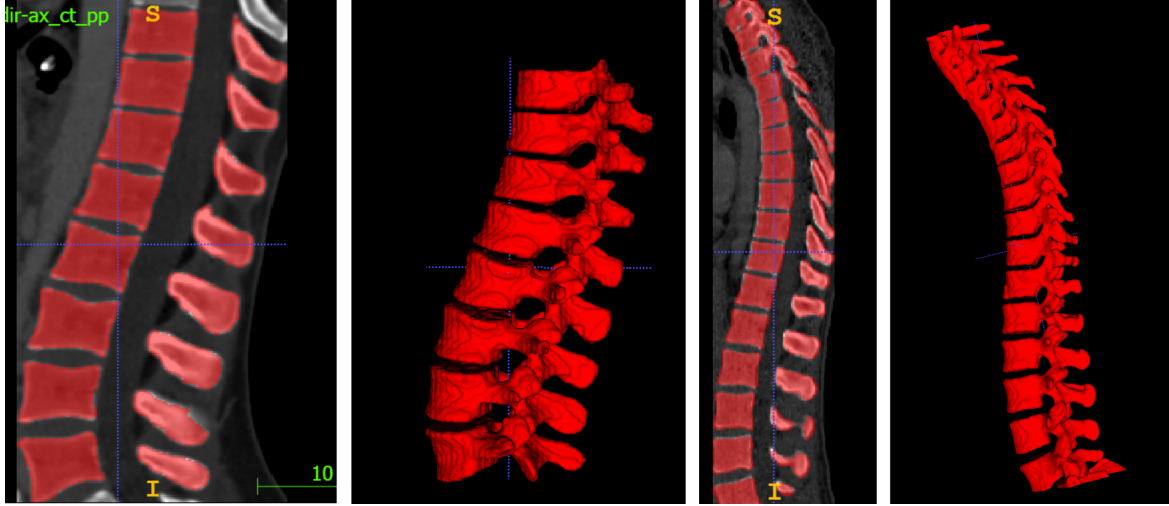


Fig. 3.13 Examples of binary spine segmentation.

3.2.4 Summary

Several methods have been proposed in the literature for the purpose of vertebrae segmentation and labeling. Many of these address only the binary spine segmentation, as in Kim et al. [270] and Vania et al. [263]. However, the methods which involve individual vertebrae segmentation are characterized by complex architectures and require tons of labeled data to be correctly implemented, as for the top scoring architectures in VerSe challenges, e.g., Payer et al. [276] and Lessman et al. [269].

The method for vertebrae labeling proposed in this thesis can be considered a novel approach, implementing k -means clustering for separating vertebral arches from bodies and k -NN classification in the context of vertebrae labeling. It is a simple yet effective solution and, most importantly, it does not require a specific training procedure. Therefore, it can also be performed without having masks provided by domain experts.

The developed algorithm addresses the issues of vertebrae identification and segmentation, which are two essential steps in understanding spine imaging data. Vertebrae segmentation is a challenging and time-consuming task, due to the size of the problem. The proposed work provides accurate results in a fast and straightforward way.

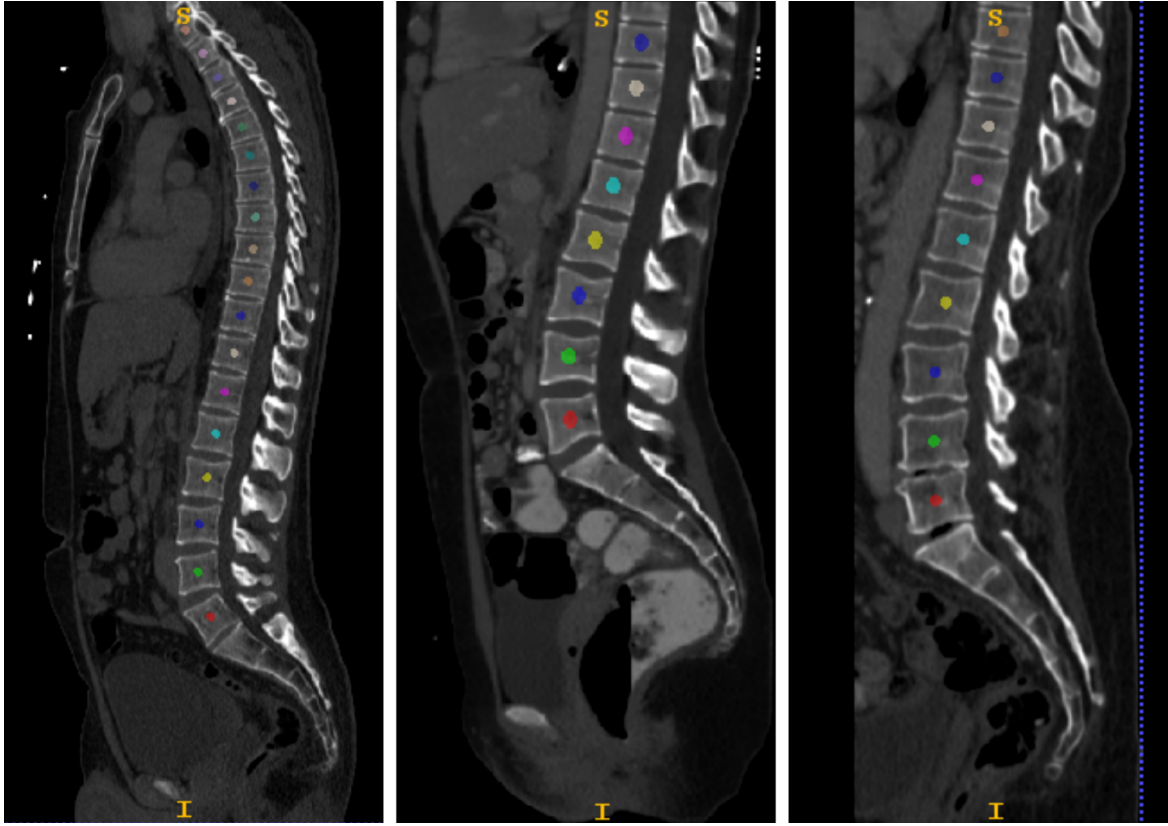


Fig. 3.14 Samples of the final output of the centroids' prediction.

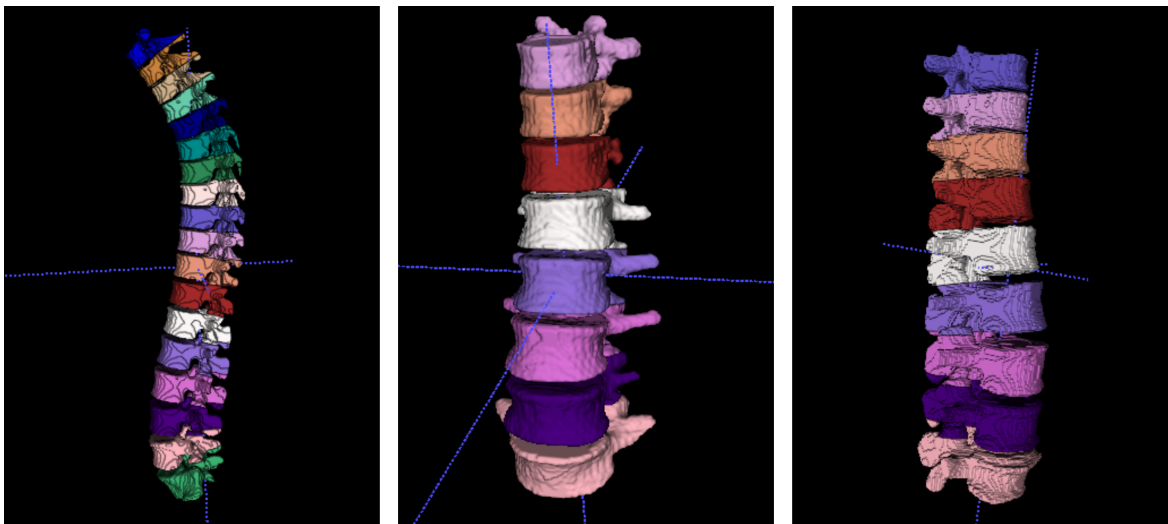
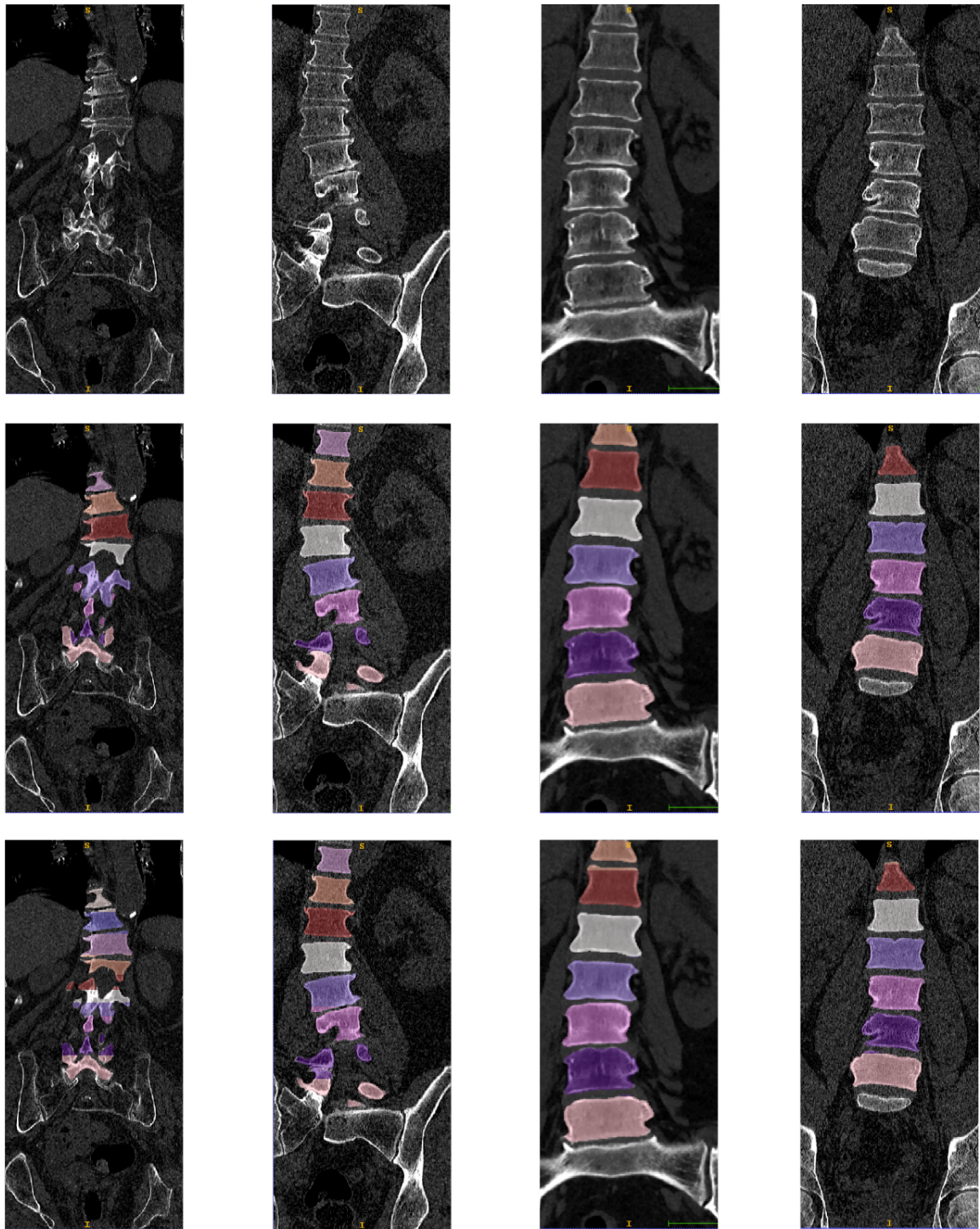


Fig. 3.15 Sample outputs of meshes of multi-class vertebrae segmentation. These meshes have been obtained exploiting ITK-Snap (<http://www.itksnap.org/>, last accessed: 6 June 2021).



a) Severe b) Severe c) Moderate d) Mild

Fig. 3.16 Comparison between Payer et al. and the proposed method for the vertebrae labeling task. The first row shows original CT scans, the second row shows predictions from Payer et al. [276], and the third row shows predictions with the proposed vertebrae labeling method. Columns: (a,b) are of patients with severe scoliosis, (c) of a patient with moderate scoliosis and (d) of a patient with mild scoliosis. It is important to note that for case (a) the proposed method did not provide an accurate result.

The clinical implications of this study also include the possibility of improving the functionalities of surgical navigators for minimally invasive spine procedures. This can help spine surgeons to operate even in unideal conditions, such as with restricted field-of-views. Moreover, the proposed method can be exploited to pre-label larger CT scan datasets with individual vertebrae annotations, so that purely supervised approaches can be enhanced. The *DSC* obtained is quite good, also if it has to be noted that the proposed approach is less general than those involved in the VerSe challenges, which covered all kinds of orientation, spacing, and field-of-view for CT spine imaging data.

3.3 Prostate Fusion Biopsy

This section of the thesis addresses the problem of providing a fast and reliable semiautomatic method for prostate segmentation from TRUS images, that can be exploited without having acquired a specific dataset for the transducer in consideration. The approach, like other existing works, is based on the theory of deformable superellipses. Two kinds of methodologies can be exploited for achieving segmentation with superellipses. In the first, image characteristics, such as edge maps or region energy, are employed for performing automatic image segmentation [294, 295]. In the second case, the geometry of the prostate is inferred from user-defined points [293].

The method of Gong et al. [294] requires the generation of edge maps in order to produce the final segmentation, whereas the approach of Saroul et al. [295] needs to make assumptions about region energy. On the other hand, the proposed approach does not need to make these assumptions on input data, given that image quality can widely vary among different transducers.

Compared to the approach of Mahdavi et al. [293], the proposed method does not require selecting points in a refined geometric way, but the user has flexibility on how to place them. Moreover, the proposed formulation allows to model also more irregular 3D shapes, since the user can insert points in an arbitrary number of slices, thus eventually obtaining shapes that go beyond tapered and warped ellipsoids.

Though, extensive experiments have been carried out to outline the best guidelines that a human operator should take into consideration when using the proposed algorithm for performing a procedure in order to minimize the number of points required and maximize the segmentation accuracy. In any case, carrying out a second iteration can mitigate eventual problems that arise after the suboptimal placing of points in the first iteration.

Lastly, an application of the proposed method in an image fusion setup with MRI is shown. The segmentation module for the MRI relies on the nnU-Net framework [258]. Segmentation masks from both domains are then registered to attain the image fusion task.

3.3.1 Materials

Fedorov et al. made publicly available a dataset containing anonymized imaging data of the human prostate of $N=3$ patients [368]. This dataset will be referred to as ZENODO throughout this section. For each patient, both MRI and TRUS examinations have been acquired. The former serves the purpose of staging the disease and the latter of allowing volumetric examination for preparing brachytherapy implant. Both modalities are 3D scalar images. Annotations provided by Fedorov et al. include manual segmentation of the whole prostate gland for both MRI and TRUS, and fiducial points placed in specific anatomical sites to improve subsequent image registration and fusion. In detail, fiducials are placed at urethra entry into the prostate at the base (UB), verumontanum (VM), and urethra entry into the prostate at the apex (UA).

In order to validate DL models for the task of MRI prostate segmentation, also the datasets PROMISE12 [290] and SAML [369, 370] have been included in the analysis.

The ZENODO dataset has been exploited to test the proposed method for TRUS segmentation, MRI segmentation, and TRUS-MRI registration, whereas PROMISE12 and SAML have been employed to validate the nnU-Net model for MRI segmentation.

Sample images for both domains, TRUS and MRI, are reported in Figure 3.17. A summarized table for the considered materials is provided in Table 3.11.

Table 3.11 *Prostate gland datasets description*. As imaging modalities, only ZENODO contains TRUS images, even in a very limited quantity, being only 3. Fiducial points are also present only in the last dataset. Seg stands for Segmentation, Reg for Registration.

Dataset	Imaging modality	Task	Number of images	Ground Truth segmentation	Fiducial points	File format
PROMISE12 [290]	MR (T2W)	MR Seg	50	✓	✗	NIFTI
SAML [369, 370]	MR (T2W)	MR Seg	116	✓	✗	NIFTI
ZENODO [368]	MR (T2W), TRUS	TRUS Seg, MR/TRUS Reg	3, 3	✓	✓	NRRD

3.3.2 Workflow

The workflow employed for achieving image fusion, starting with segmentation for both imaging modalities, namely TRUS and MRI, is reported in Figure 3.18. In clinical practice,

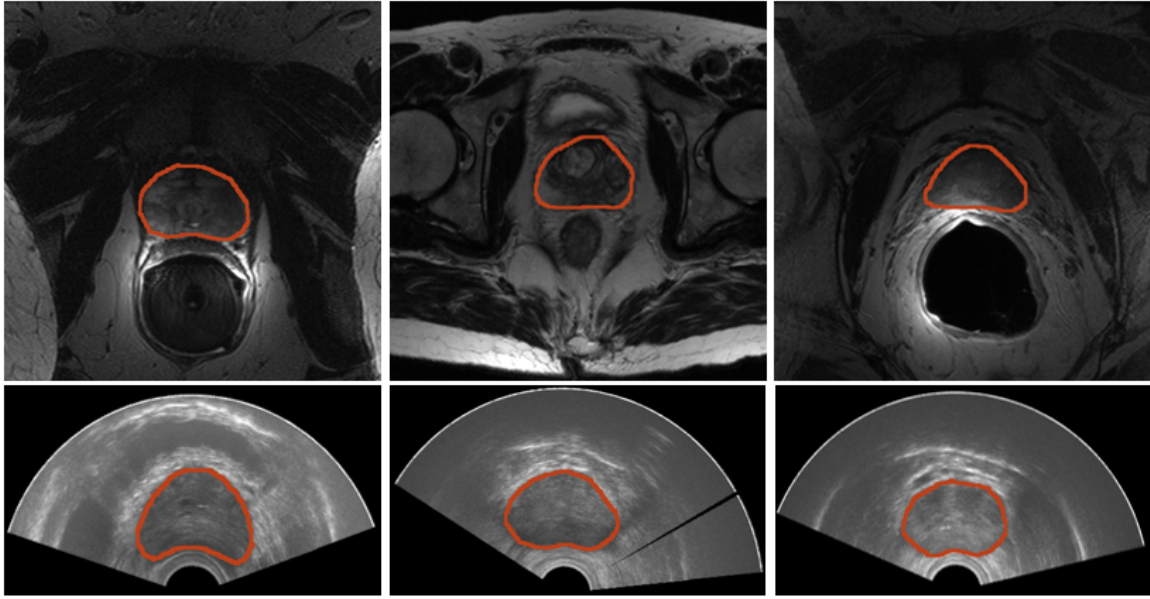


Fig. 3.17 *Samples of images from MRI and TRUS modalities.* (Top) Prostate MRI. From left to right, a sample image for each of the datasets PROMISE12, SAML, and ZENODO is shown. (Bottom) Three sample prostate TRUS from the ZENODO dataset.

segmentation does not happen at the same time, since MRI segmentation can be achieved preoperatively, whereas TRUS segmentation has to be obtained intraoperatively, at the start of the prostate biopsy procedure.

Segmentation from MRI involves a pre-processing stage so that images can be fed to a deep learning architecture, the nnU-Net. Lastly, post-processing is performed, with the aim to remove noisy elements from images (e.g., only one connected component is expected), increasing segmentation accuracy. The described operations can be carried out in a fully automatic way. MR images are especially important for identifying the target region for biopsy since they have better contrast than other imaging modalities. Details are described in Section 3.3.2.1.

Segmentation from TRUS is achieved with a semiautomatic algorithm, which requires input points from the user. The physician has to annotate points in at least three slices of the prostate gland in axial planes, taking care when placing points in the deformed zones (the transducer itself introduces deformation). Starting from this point, a deformable superellipse is fitted with an optimization algorithm. Then, interpolation is employed to achieve the 3D reconstruction of the prostate gland. The entire procedure is explained in Section 3.3.2.2.

Then, with both segmentation masks from TRUS and MRI modalities available, registration can be performed, enabling image fusion, which allows tissue coming from both

modalities to be seen at the same time. Optionally, a set of anatomical landmarks can be inserted by the user to ease and constrain the registration optimization step. The procedure is presented in Section 3.3.2.3.

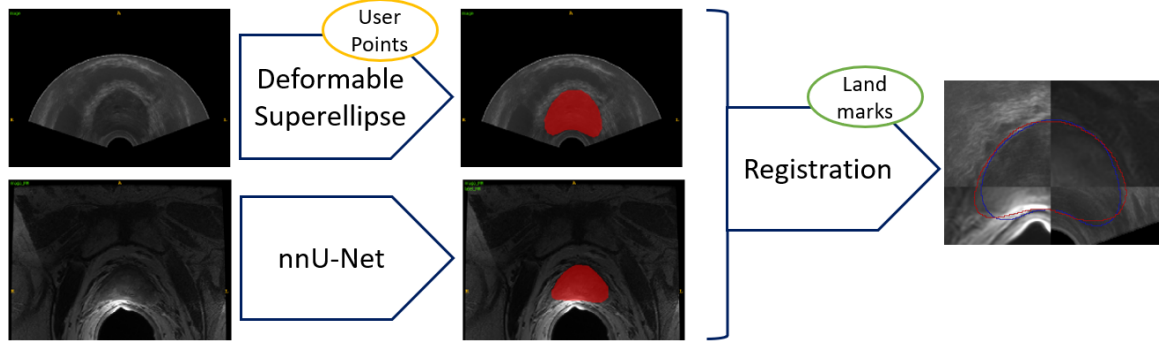


Fig. 3.18 *Workflow for TRUS and MRI segmentation and subsequent image fusion.* Segmentation from TRUS is achieved in a semiautomatic way by fitting a 3D model based on deformable superellipses starting from user-defined points in at least three slices. Segmentation from MRI has been performed fully automatically by exploiting the nnU-Net framework. Registration can be either performed in an automatic way, or the user can add anatomical landmarks to constrain the space of transformations.

3.3.2.1 MRI Segmentation

The semantic segmentation of the prostate gland from MRI can be efficiently met via DL techniques, as fully convolutional neural networks [16]. Semantic segmentation, which poses the basis for subsequent classification and characterization tasks [31, 371], is essential in numerous clinical applications including artificial intelligence in diagnostic support systems, therapy planning, intraoperative assistance, and monitoring of tumor growth.

As introduced in Section 2.1.2, semantic segmentation is a Computer Vision task that can be computed with DL algorithms and consists in labeling each pixel of an input image, without recognizing the different instances of objects [18, 19]; it is possible to see semantic segmentation as a problem of conversion from image to image, where the input image is the original image and each pixel intensity value of the output image indicates the relation of that pixel to the associated class [89].

Most semantic segmentation architectures are based on encoder-decoder networks. The process of feature extraction or sub-sampling is carried out by the encoder. Decoding is an up-sampling operation, in which the spatial information output from the encoding layer is reconstructed, increasing the spatial resolution. The encoder-decoder structures have been implemented in different convolutional network architectures, including SegNet [70], U-Net

[72], U-Net 3D [73] and V-Net [71]. Besides prostate segmentation, applications in medical imaging tasks of those architectures encompass liver vessels delineation [56], segments classification [35], lung COVID-19 lesions segmentation [37], and vertebrae segmentation [36].

In the work presented in this section, to perform the semantic segmentation of the prostate gland from MRI, the nnU-Net framework has been exploited. In this way, semantic segmentation tasks can be tackled with standardized pipelines [84, 258]. The employed architecture is based on those of U-Net and U-Net 3D. The nnU-Net framework was originally conceived during the Medical Decathlon Segmentation Challenge [372], where it emerged as the leading approach in all tasks. The advantages of this method consist of automatic configurations of pre-processing, data augmentation, training, inference, and post-processing. Parameters to set for training nnU-Net include the number of epochs, initial learning rate, batch size, patch size, and the combination of dice loss and cross-entropy to implement as the model loss function.

3.3.2.2 TRUS Segmentation

Segmentation of anatomical structures in noisy data, such as TRUS images, is a complex task since boundaries are not clearly defined, as shown in Figure 2.6. Therefore, the adoption of prior information about the geometric structure of interest is useful to constrain the model deformation [373, 374]. Deformable models can be used to achieve this result.

Geometry, physics, and mathematical optimization lie the foundation for the segmentation algorithms based on deformable models [373]. The constraint on the model shape is derived from geometry, the evolution of the shape in space is governed by physical theories, and the operation of fitting the model to the accessible data is made possible by optimization theory [375]. Segmentation of anatomical structures in deformable models is achieved by exploiting an energy minimization framework. Two kinds of mathematical terms are considered: internal and external energies. The deformable model is propagated in the direction of the object contours by external energies, whereas the smoothness of the boundaries is preserved by internal energies.

The deformable model framework includes various methodologies, such as deformable mesh, active shape models, level sets, active contour models, and curve fitting [280]. More advanced approaches may include a mixture of these techniques, with the idea that merging information concerning boundaries known *a priori*, region, shape, and features of the prostate region can provide more accurate models, like the deformable superellipse formulation of Gong et al. [294].

Shape Models In a wide variety of medical imaging scenarios, the general location, orientation, and shape of the objects of interest are known *a priori*. As reported by previous studies concerning TRUS images, prostate contours appear smooth and with a closed-near convex shape [294]. This information can be embedded into the deformable model in different forms: initial conditions, way of constraining model shape parameters, and the procedure for model fitting. Global shape properties can be modeled with parametric shape models. The advantage of this technique is not requiring the presence of anatomical landmarks.

Furthermore, representation of the shapes can be tackled with many different methods [376, 377]. For instance, Tutar et al. [378] proposed to model the 3D prostate boundaries with spherical harmonics of degree eight. Local deformations can be controlled, thanks to the exploitation of parameters, leading to the capacity of modeling complex shapes. On the other side, there is an increment in the computational complexity.

Similarly, reducing the number and range of parameters can allow modeling the global shape in approaches that are stable and rapid from a numerical perspective, leading to more compact representations. In the following section, the deformable superellipse, a powerful model for the geometry of the prostate gland [294], is introduced. When the deformable superellipse is not capable of properly capturing all the nuances of the prostate region in a 2D slice, bidimensional B-splines [379] can be exploited in the proposed approach, obtaining very refined results but not losing the possibility to model a regular 3D shape with a relatively low number of parameters.

Deformable Superellipses Superellipses consent to obtain a natural generalization of the ellipses' shapes. Different base geometrical shapes can be modeled through superellipses, such as ellipses, parallelograms, rectangles, and pinched diamonds by handling a small number of parameters [380, 381]. Examples of shapes that can be modeled by superellipses are portrayed in Figure 3.19. The straightforward 3D generalization of the superellipse, the superellipsoid, has not been considered since it makes assumptions about the 3D regularity of the prostate shape which are too simplistic.

A centered superellipse can be expressed in the following **parametric** form, as reported in Equation (3.15):

$$\begin{cases} x = a_x \cdot |\cos(\theta)|^{\frac{2}{\varepsilon}} \cdot \text{sign}(\cos(\theta)) \\ y = a_y \cdot |\sin(\theta)|^{\frac{2}{\varepsilon}} \cdot \text{sign}(\sin(\theta)) \end{cases} \quad (3.15)$$

where the size parameters $a_x > 0$, $a_y > 0$ define the length of the semi axes, and $\varepsilon > 0$ specifies the squareness in 2D plane, as shown in Figure 3.19. The corresponding **implicit**

form is given by the Equation (3.16):

$$\left| \frac{x}{a_x} \right|^\varepsilon + \left| \frac{y}{a_y} \right|^\varepsilon = 1 \quad (3.16)$$

The *inside-outside* function is reported in Equation (3.17):

$$f(x,y) = \left| \frac{x}{a_x} \right|^\varepsilon + \left| \frac{y}{a_y} \right|^\varepsilon \quad (3.17)$$

where: if $f(x,y) = 1$, then the point (x,y) belongs to the superellipse; if $f(x,y) > 1$, then the point (x,y) lies outside the superellipse; if $f(x,y) < 1$, then the point (x,y) lies inside the superellipse.

The superellipse model does not permit, in this version, modeling all deformations which are required to build a proper representation of the prostate gland. Nonetheless, geometric deformations, such as translation, rotation, tapering, and bending, can result in a broad range of shapes that are modeled by the deformable superellipse [382, 383]. Moreover, these transformations can be modeled with a few parameters, given that translation with respect to an axis, rotation, tapering, and bending are described each with a single parameter. Deformable superellipse can then be characterized by a parameter vector $\mathbf{p}$, defined as in Equation (3.18):

$$\mathbf{p} = \{a_x, a_y, l_x, l_y, r, \varepsilon, t, b\} \quad (3.18)$$

where ε is the squareness parameter and a_x, a_y are the semi-axes lengths defined above. Other parameters are those involved in the global similarity transformations for superellipses [294]: l_x, l_y are the translations along x and y axes, r is the rotation angle, t and b model the tapering and circular bending on the y -axis, respectively.

Details of all these geometric transformations are reported in Paragraph 3.3.2.2-**Geometric Transformations**, whereas inverse transformations are reported in Paragraph 3.3.2.2-**Inverse Transformations**. Examples of deformable superellipse modeled by variations in tapering and bending are reported in Figure 3.19.

Geometric Transformations Translation (l_x, l_y):

$$\begin{cases} x' = x + l_x \\ y' = y + l_y \end{cases} \quad (3.19)$$

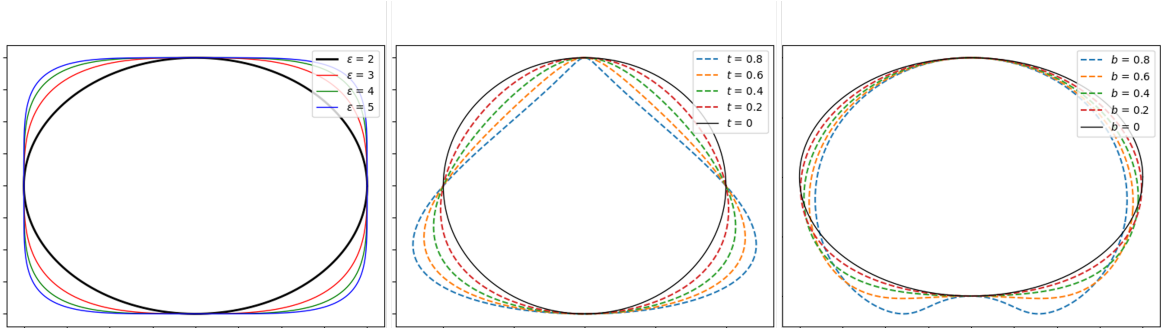


Fig. 3.19 *Deformable superellipse modeling capabilities examples*. The left image represents the superellipse varying squareness ε parameter, whereas the middle one depicts the deformable superellipse varying tapering t parameter, and the right one portrays the deformable superellipse varying circular bending b parameter.

Rotation (r):

$$\begin{cases} x' = x \cdot \cos(r) - y \cdot \sin(r) \\ y' = x \cdot \sin(r) + y \cdot \cos(r) \end{cases} \quad (3.20)$$

Linear tapering along y-axis (t):

$$\begin{cases} x' = x \cdot \left(\frac{t \cdot y}{a_y} + 1 \right) \\ y' = y \end{cases} \quad (3.21)$$

Circular bending along the y-axis (b):

$$\begin{cases} x' = \left(\frac{a_y}{b} - y \right) \cdot \sin\left(\frac{x}{\frac{a_y}{b} - y} \right) \\ y' = \frac{a_y}{b} - \left(\frac{a_y}{b} - y \right) \cdot \cos\left(\frac{x}{\frac{a_y}{b} - y} \right) \end{cases} \quad (3.22)$$

Inverse Transformations Inverse translation (l_x, l_y):

$$\begin{cases} x = x' - l_x \\ y = y' - l_y \end{cases} \quad (3.23)$$

Inverse rotation (r):

$$\begin{cases} x = +x' \cdot \cos(r) + y' \cdot \sin(r) \\ y = -x' \cdot \sin(r) + y' \cdot \cos(r) \end{cases} \quad (3.24)$$

Inverse linear tapering along y-axis (t):

$$\begin{cases} x = \frac{x'}{\frac{t \cdot y}{a_y} + 1} \\ y = y' \end{cases} \quad (3.25)$$

Inverse circular bending along the y-axis (b):

$$\begin{cases} x = -\text{sign}(b) \cdot \arctan\left(\frac{x'}{y' - c}\right) \cdot \sqrt{(x')^2 + (y' - \frac{a_y}{b})^2} \\ y = \frac{a_y}{b} - \text{sign}(b) \cdot \sqrt{(x')^2 + (y' - \frac{a_y}{b})^2} \end{cases} \quad (3.26)$$

Optimization Framework In the Bayesian framework proposed in Gong et al. [294], the authors assumed that some parameters (those concerning shape) have a Gaussian distribution as prior, $N(\mathbf{p}_s)$, whereas others (those concerning pose) have a Uniform distribution as prior, $U(\mathbf{p}_p)$. The edge strength likelihood is denoted as E . Then, according to the Bayes rule, the posterior probability can be modeled as in Equation (3.27):

$$Pr(\mathbf{p} | E) = \frac{Pr(\mathbf{p}) \cdot Pr(E | \mathbf{p})}{Pr(E)} = \frac{Un(\mathbf{p}_p) \cdot N(\mathbf{p}_s) \cdot Pr(E | \mathbf{p})}{Pr(E)} \propto Un(\mathbf{p}_p) \cdot N(\mathbf{p}_s) \cdot Pr(E | \mathbf{p}) \quad (3.27)$$

This results in optimizing the log-likelihood in Equation (3.28):

$$L = \ln(Pr(\mathbf{p}_s)) + \ln(Pr(E|\mathbf{p})) \quad (3.28)$$

Proposed Approach In the proposed approach, the deformable superellipse is modeled as specified in Section 3.3.2.2-**Deformable Superellipses**. Geometric deformations to the fundamental superellipse shape can be obtained as reported in Section 3.3.2.2-**Geometric Transformations**.

The problem of modeling $Pr(\mathbf{p} | E)$, as in Equation (3.27), is that it requires to have prior data on edge maps from images of the same kind of those obtained with the ultrasound device that will be used for carrying out the procedures. When it is not feasible to collect such images in advance, it may be preferable to model $Pr(\mathbf{p} | U)$, where U is a set of user-defined points. If the model does not need to make rigid assumptions about U , it can provide a fast and reliable system for achieving prostate gland segmentation with only moderate user interaction and without the need to build a large training set.

Therefore, in the proposed formulation, the posterior probability can be written as reported in Equation (3.29):

$$Pr(\mathbf{p} | U) = \frac{Pr(\mathbf{p}) \cdot Pr(U | \mathbf{p})}{Pr(U)} \propto Un(\mathbf{p}_p) \cdot N(\mathbf{p}_s) \cdot Pr(U | \mathbf{p}) \quad (3.29)$$

The prior about shape parameters can be optimized by maximizing Equation (3.30) [294]:

$$\ln(Pr(\mathbf{p}_s)) = - \sum_j \left[\frac{(p_j - m_j)^2}{2 \cdot \sigma_j^2} \right] \quad (3.30)$$

Instead, the likelihood linked to the term $Pr(U | \mathbf{p})$ can be maximized by optimizing the energy in Equation (3.31), where U is the set of user-defined points, C is the polygon representing the prostate mask boundaries, d is the point-to-polygon distance, and $E(C; U)$ is the energy function to minimize.

$$E(C; U) = \sum_{(x,y) \in U} d(C, (x,y))^2 \quad (3.31)$$

A 3D model can be reconstructed by performing linear interpolation of the parameters involved in the vector $\mathbf{p}$, after that 2D superellipses have been fit to the slices where the user has inserted points. To build a 3D model of the prostate gland, a minimum of three slices have to be labeled. The annotated slices must include the base, apex, and mid-gland regions of the prostate gland. On the base and apex, a minimum of 4 points must be inserted by the user, whereas on the mid-gland a minimum of 6 is recommended. For mid-gland cases which have irregular shapes, a number of points up to 12 may be beneficial.

Since the user can add more than three slices, shapes that are more complex than one tapered and warped superellipsoid or two semi-superellipsoids can be obtained. The following paragraph describes how the optimization procedure is executed when an operator is involved.

Implementation Details The general workflow employed for TRUS segmentation is reported in Figure 3.20.

First, the user is asked to select points from at least three slices of the TRUS volume. In every slice, the user has to select a variety of points ranging approximately from 4 to 12, as detailed at the end of Section 3.3.2.2. In order to ease this process for the experiments realized during this research, a JSON interface with the popular 3D Slicer software [384] has been realized for this research work.

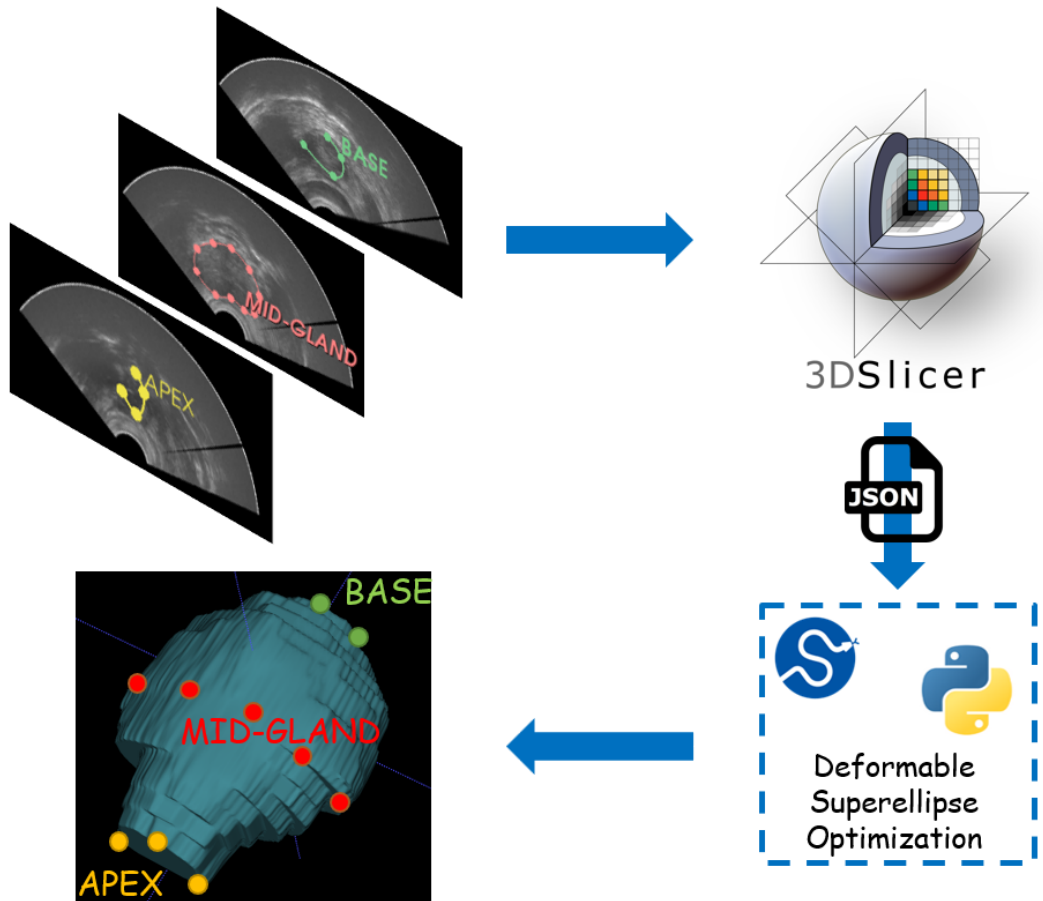


Fig. 3.20 *Workflow for TRUS segmentation with the developed methodology.* The operator needs to annotate some points in at least the apex, base, and mid-gland of the prostate. Then, a JSON file is fed as input to an optimization routine that fits the best 2D superellipse in every annotated slice. Then, a 3D model is built by linearly interpolating 2D models. 3D Slicer has been exploited as GUI to speed up and ease the process.

The user can enter two types of models when inserting points. The first is the superellipse, whereas the second exploits bidimensional B-splines (as implemented by the method *scipy.interpolate.splprep*). For the purposes of 3D modeling, a superellipse is then fitted to the spline in the second case. For mid-gland slices, the B-spline configuration, especially when 10-12 points are annotated by the user, is the recommended way to proceed. When there are few annotated points, deformable superellipse is more likely to properly work, considering that it has a relatively low number of parameters. In particular, in the configuration with the least possible number of points, where the user places 4 points at the base, 6 points at mid-gland, and 4 points at the apex, the deformable superellipse should be exploited.

In order to effectively implement the optimization procedure of the 2D superellipse to the slice points, an iterative minimization procedure has been carried out. At every iteration, the optimizer passes a vector $\mathbf{p}$ of parameters to a superellipse class, which has the twofold purpose to (i) build an object with the given parameters, (ii) measure its energy with respect to the user-defined points.

After the object is created, the *inside-outside* function reported in Equation (3.17) is used to create a mask of points that satisfies the condition for the centered superellipse. Then, these points are transformed by using deformations in the following order: rotation, as defined in Equation (3.20); linear tapering along y-axis, as defined in Equation (3.21); circular bending along y-axis, as defined in Equation (3.22); translation, as defined in Equation (3.19).

The mask obtained by these transformations is subject to the morphological closing operator since holes arise during the transformation process. Then, energy for the built superellipse can be defined as the sum of distances from user-defined points to the polygon of the mask boundary. The point-to-polygon distance can be calculated with the *pointPolygonTest* method from the OpenCV library. At the end of the minimization procedure, the optimizer will find the best vector $\mathbf{p}$ for the input points given by the user.

Lastly, the 2D Deformable Superellipse models fit in multiple slices (at least three including base, apex, and mid-gland) are exploited to reconstruct the 3D volume by performing linear interpolation of the parameters contained in the vector $\mathbf{p}$. The program also returns a list of JSON files which can be loaded in 3D Slicer to refine the segmentation results and eventually perform a second iteration. In the second user iteration, B-splines are exploited for providing the contour of the prostate gland, since the user only needs to adjust boundary points provided by the previous iteration of the algorithm.

3.3.2.3 MRI-TRUS Registration

The described registration algorithm is segmentation-based. Indeed, both MRI and TRUS segmentation masks are required for performing the procedure. Other authors considered this step fundamental too [385, 386]. The particular challenge of MRI-TRUS registration is that the anatomical areas visible in one modality may not be visible in the other.

Before starting with the registration procedure, pre-processing has been performed with the purpose to improve and ease the fusion algorithm results. First, the 3D images have been cropped into 3D bounding boxes (i.e. Volume of Interest (VOI)) that extend for 10mm over the margin delineated by the segmentation mask. Then, the VOIs have been resampled to make them isotropic and with the same resolution for both modalities.

For the binary segmentation mask, the Nearest Neighbour interpolator has been employed to perform the resampling. Output resolution has been set to $0.3\text{mm} \times 0.3\text{mm} \times 0.3\text{mm}$. Segmentation masks have been smoothed with a gaussian kernel with $\sigma = 3$ [386]. Lastly, the Maurer signed distance transformation, which exploits the Euclidean Distance transform [387], has been applied to the segmentation masks. The steps involved in the registration pre-processing are reported in Figure 3.21.

The purpose of the initialization is to simplify the calculation of the center of rotation and translation needed for the rigid transformation. Two kinds of initialization have been considered: (i) based on the center of images; (ii) based on a set of landmarks.

In the first case, centers of images are calculated in the coordinate spatial system considering the origin, dimensions, and spacing of images. The geometric center of the moving image is given as the initial center of the rigid transformation, and the vector that goes from the center of the fixed image to the center of the moving image is given as the initial translation vector.

The second approach, instead, determines an initial transformation by considering a set of landmarks. It determines the optimal transform that can map the fixed image and the moving image with respect to the least square errors of the levels of intensity [388].

Since the proposed approach aims to perform the registration of distance maps whose intensity values have the same range of values and meaning, the dissimilarity measure employed is the sum of squares of intensity differences (SSD). Lower values of the said metric correspond to better results. The optimizer employed is based on gradient descent, and is targeted at finding the set of parameters that define a transformation that optimizes the metric as better as possible. The overall workflow employed for the registration with all the various components described in this section is portrayed in Figure 3.22.

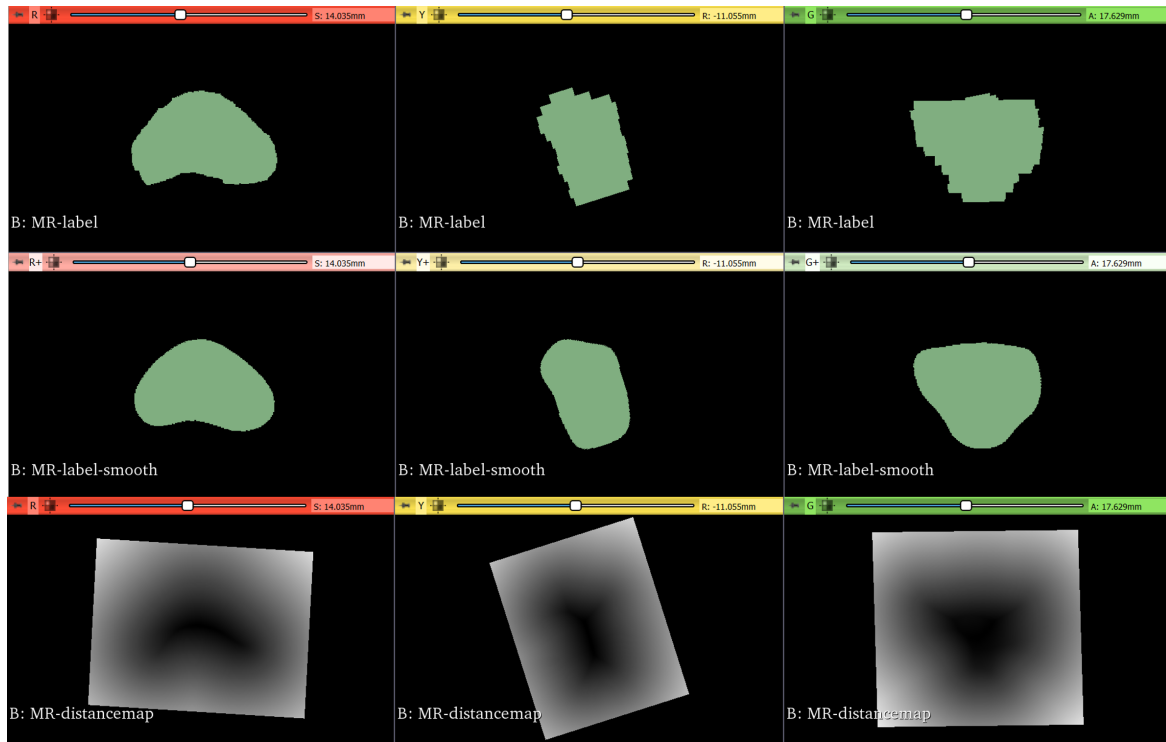


Fig. 3.21 *Image registration pre-processing*. Original ground truth segmentation masks are reported in the top row. Then, they are smoothed with a gaussian filter, as depicted in the middle row. Lastly, Distance Maps are obtained from the smoothed masks, as shown in the bottom row.

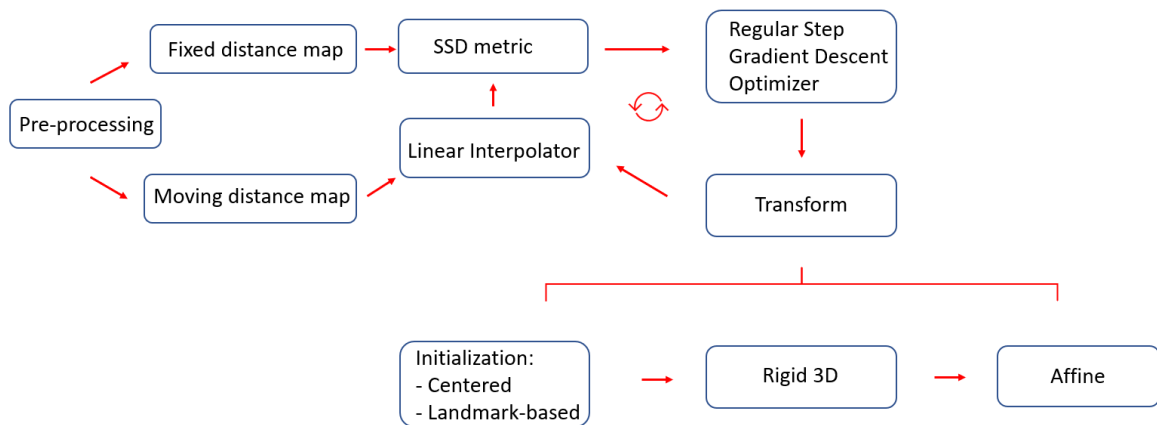


Fig. 3.22 *Proposed registration workflow*. It starts with pre-processing segmentation masks, to make them isotropic and at the same resolution. Thereafter, SSD is exploited as the metric to perform the registration, whereas gradient descent is adopted as the optimizer. Two kinds of initialization have been considered: one based on centers and the other based on landmarks.

3.3.3 Results

3.3.3.1 Segmentation

MRI Quantitative results for MRI segmentation with nnU-Net are reported in Table 3.12, whereas qualitative results, as segmentation masks, are depicted in Figure 3.23. The nnU-Net model trained on the PROMISE12 challenge has been used for obtaining the best results [258, 290]. The SAML-V dataset has been obtained by sampling 24 images for validation from the SAML dataset. It is worth noting that the Dice coefficient is higher than 88 % and ASSD is less than 1 mm for both validation sets under consideration, showing the reliability of the nnU-Net framework for automatic MRI segmentation of the prostate region.

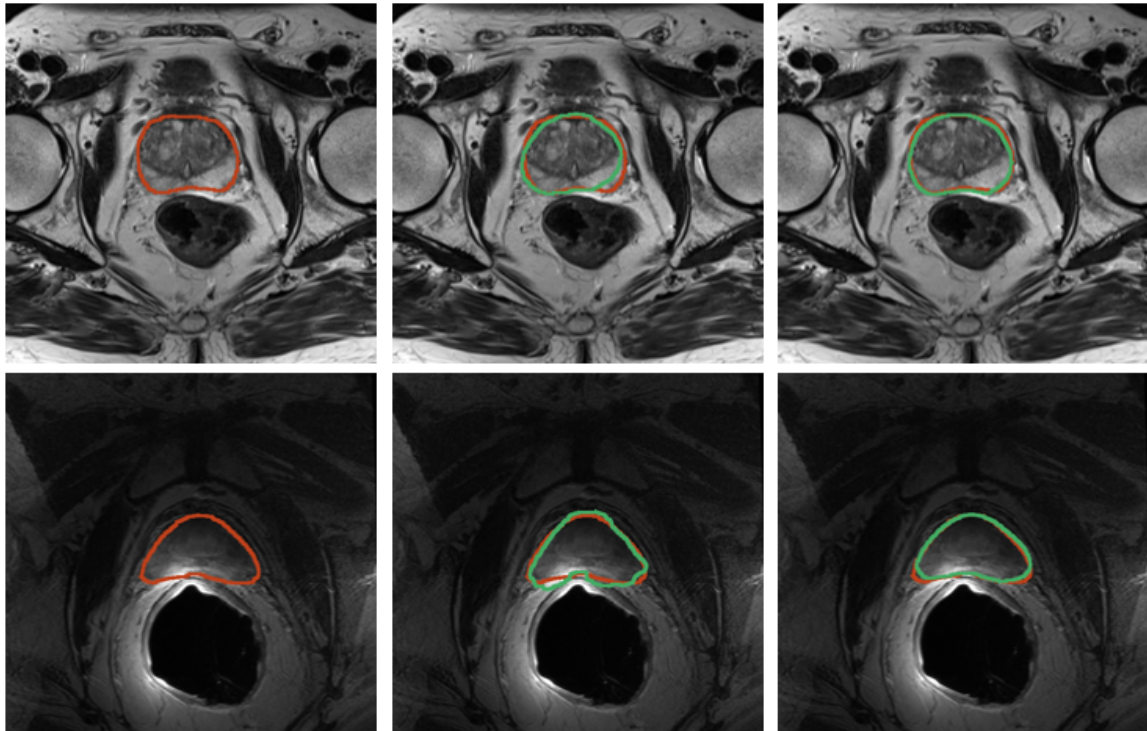


Fig. 3.23 *Results for prostate segmentation from MRI.* The top row contains images from the SAML dataset, whereas the second row encloses slices from ZENODO. The ground truth is represented in red, whereas the predictions from the nnU-Net models are colored in green. The middle image shows the prediction mask for the nnU-Net trained for only 10 epochs, whereas the right image depicts the prediction mask for the one trained on the PROMISE12 dataset.

TRUS Quantitative results for TRUS segmentation with the developed methodology based on deformable superellipses are reported in Table 3.13, whereas sample segmentation images

Table 3.12 *Quantitative metrics results for MRI prostate segmentation with nnU-Net*. Performance has been measured on two validation sets.

Train Set	Epochs	Test Set	Dice [%]	RVD [%]	HD [mm]	ASSD [mm]
PROMISE12	1000	SAML-V	88.18 ± 10.53	17.58 ± 31.61	21.03 ± 51.06	0.86 ± 1.14
PROMISE12	1000	ZENODO	91.17 ± 1.19	4.13 ± 8.79	16.11 ± 3.56	0.26 ± 0.01

are depicted in Figure 3.24. Three experiments have been conducted for each case, placing 4 points on the base and 4 on the apex, using only the superellipse to fit the contours. Instead, on the mid-gland, a number of points varying from 10 to 12 has been considered, exploiting B-splines before fitting the superellipse to finally achieve the 3D modeling of the prostate gland. Results are reported as mean $\pm$ std of the experiments done on each case. It is possible to see that results are overall considerable, being the Dice coefficient greater than 87% in all cases. Moreover, the proposed implementation is iterative, so that the user can refine the results until it reaches the desired performance. For the purposes of this research, the experiments stopped at the second iteration, which allowed the enhancement of results in all cases.

Table 3.13 *Quantitative metrics results for TRUS prostate segmentation with the proposed superellipse-based approach*. Results are reported for both the 1st and the 2nd iterations of the algorithm execution.

Metrics		Dice [%]	RVD [%]	HD [mm]	ASSD [mm]
Case 9	1 st	87.15 ± 2.41	-13.27 ± 8.34	25.12 ± 5.58	0.53 ± 0.120
	2 nd	88.56 ± 2.66	-9.44 ± 8.88	16.10 ± 7.12	0.38 ± 0.022
Case 10	1 st	89.31 ± 1.13	-12.21 ± 3.06	9.25 ± 2.41	0.23 ± 0.020
	2 nd	92.57 ± 0.45	-4.86 ± 0.36	9.37 ± 2.53	0.17 ± 0.015
Case 12	1 st	90.76 ± 1.39	-5.46 ± 3.61	23.30 ± 9.58	0.30 ± 0.049
	2 nd	92.47 ± 0.30	-1.87 ± 1.24	21.26 ± 8.22	0.23 ± 0.048

3.3.3.2 Registration

Quantitative results for registration across the two considered imaging modalities, TRUS and MRI, are reported in Table 3.14 for the configurations with and without landmarks, respectively. An example of the workflow for the image fusion is depicted in Figure 3.25. The Dice coefficient is higher than 91 % for all the cases, and HD is less than 4 mm, demonstrating that the developed registration method is promising.

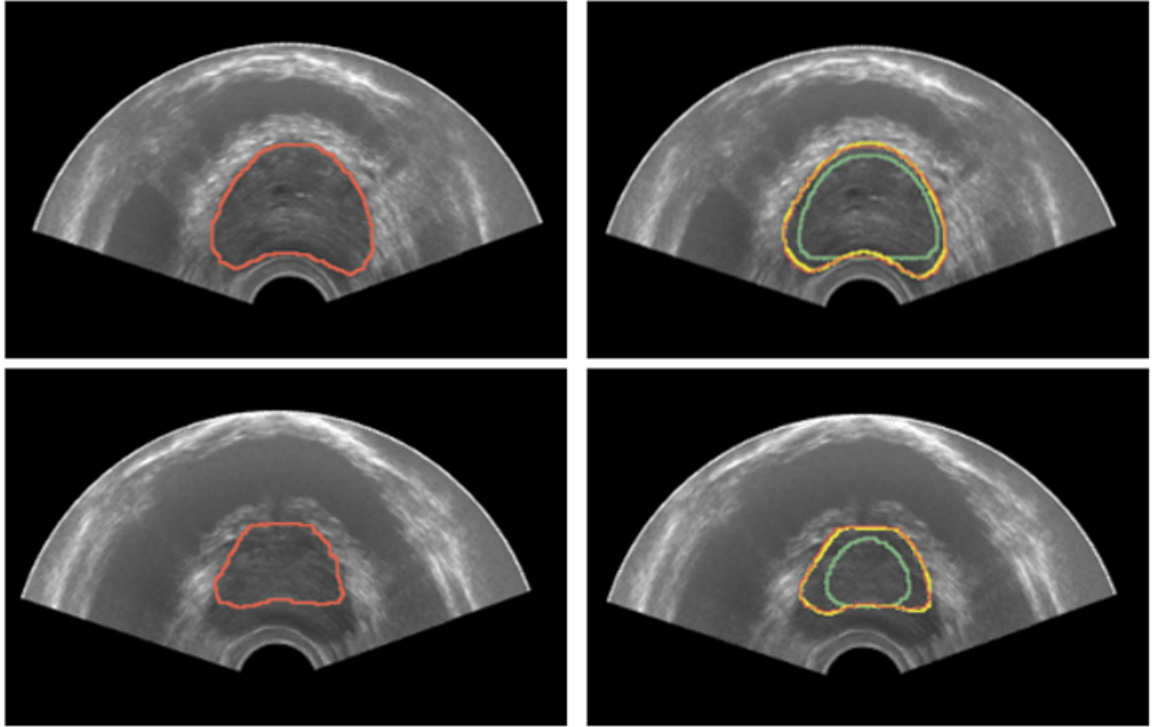


Fig. 3.24 *Qualitative results for prostate segmentation from TRUS*. The left image portrays the ground truth prostate mask in red. The right one instead depicts the segmentation results both after the first and second iteration in green and yellow colors, respectively.

Table 3.14 *Quantitative registration results*. Results are shown in two different configurations. The first exploits as the initializer the center of the images, whereas the second employs a set of landmarks.

Experiments	Dice [%]	Jaccard [%]	RVD [%]	HD [mm]
case10-center	91.77	84.79	-0.86	3.77
case10-landmarks	91.78	84.80	-0.87	3.77
case12-center	94.82	90.15	-5.79	2.12
case12-landmarks	94.85	90.21	-5.79	2.09
case9-center	93.61	87.99	-1.86	3.55
case9-landmarks	93.60	87.98	-1.88	3.60

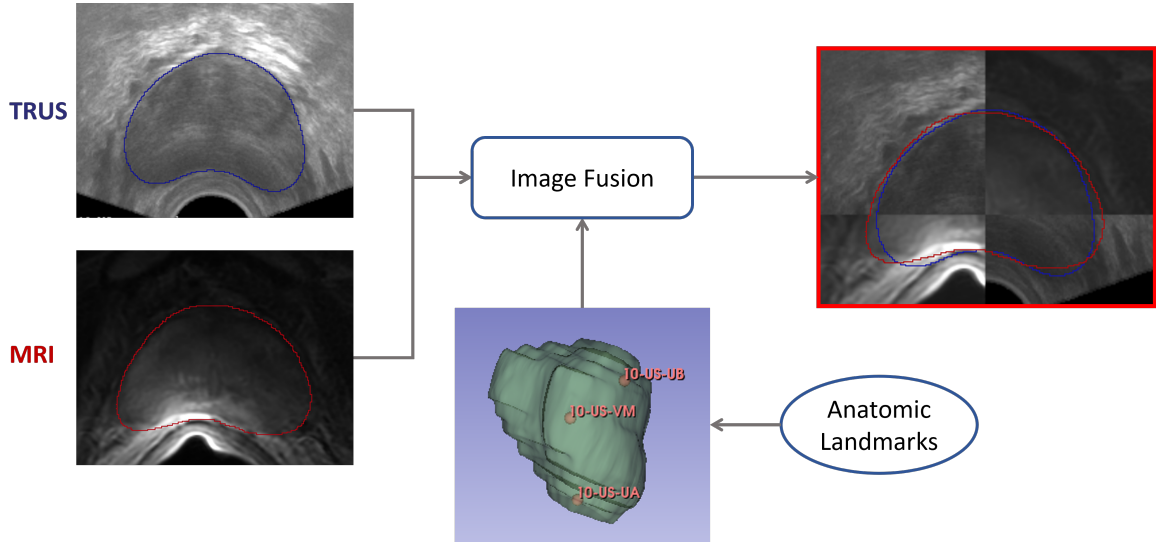


Fig. 3.25 *Image fusion workflow*. Segmentation masks are obtained for both domains: TRUS and MRI. Then, the registration is performed as described in Section 3.3.2.3, so that images can be fused. Both masks have been shown after the registration procedure.

3.3.3.3 Discussion

Prostate segmentation is a pivotal, but strenuous to accomplish, task that is required for targeted prostate biopsy procedures. Moreover, every transducer for TRUS can produce different images, resulting in a variety of conditions that makes it difficult to transfer what has been learned on one dataset to another. Lastly, the lack of annotated datasets for TRUS segmentation adds to the peculiarity of the task. Indeed, the ZENODO dataset, which consists of merely 3 images, was the only one disposable for the research work herein described.

The deformable superellipses are shape models that allow modeling a variety of geometry deformations starting from ellipses, which can resemble the most common prostate shapes. In fact, the prostate shape can be well approximated by a tapered ellipsoid [293]. When the procedure is performed, the transducer induces a slightly posterior deformation in the patient's prostate which can be modeled, for instance with the bending parameter b .

Therefore, this work proposed a novel formulation of the deformable superellipse to make it a suitable method for TRUS segmentation, also in the absence of training data from a given transducer. Other approaches, like that of Gong et al. [294], require edge detection algorithms, so that could be exploited for automatic segmentation, but on the other side, demand training data from the specific transducer. The advantage of the proposed method is that it can be applied in any circumstance, only necessitating a moderate interaction with the physician, and always yields considerable results.

In the experiments carried out for this study, the proposed method required 41 ± 7 s for placing the points in three or four slices, whereas it took 5 ± 1 s to build the 3D model. The time needed for the second iteration was more variable— 74 ± 32 s. The superellipse implementation of Mahdavi et al. [293] took 32 ± 14 s for initialization, which is similar to the time needed to place the initial user-defined points in the proposed approach. On the computational side, it needed 14 ± 1 s, which is more than the considered implementation. Furthermore, in their case, segmentation refinement can be performed by the user, with a time ranging between 1 and 3 minutes. It is not possible to directly compare the proposed approach with the work of Gong et al. [294] since their method is capable of performing segmentation in less than 5 s per slice, but it only delineates 2D boundaries.

The developed methodology managed to achieve respectable results, reaching the Dice coefficient a value higher than 87% in all images considered in the test set, composed by the ZENODO dataset. Then, the research focused on proving the applicability of this module in a targeted biopsy setup. So, the nnU-Net framework has been exploited for the task of performing segmentation from MRI, achieving a Dice coefficient of more than 88% on the SAML-V dataset and higher than 91% on the ZENODO dataset.

Lastly, a custom registration procedure has been developed, which consented to reach a Dice coefficient higher than 91% and HD lower than 4mm in all cases, showing the effectiveness of the proposed framework in a clinical application. In the registration framework, two initializers have been considered: (i) one based on centers of images and (ii) one which relied upon a set of landmarks. From the obtained results, it is possible to note that the former allowed to reach Dice coefficients of 91.77%, 94.82%, 93.61%, and HD of 3.77mm, 2.12mm, 3.55mm, whereas the latter managed to achieve Dice coefficients of 91.78%, 94.85%, 93.60%, and HD of 3.77mm, 2.09mm, 2.29mm. Hence, the two methods provide similar results. Therefore, also a simpler center-based initialization can be adopted for the affine registration procedure.

Overall, the obtained results, for both segmentation methods, are satisfactory for the implementation in a targeted prostate biopsy setup. The registration framework can eventually be improved, by exploiting deformable models also in this stage, eventually allowing better results for the image fusion procedure.

3.3.4 Summary

Prostate segmentation from MRI and TRUS is a complex challenge but may have a huge impact on clinical setups for fusion biopsy. For what concerns MRI, with the advent of the

nnU-Net framework, the challenge is more easily met since a standardized pipeline can be employed for semantic segmentation. However, there is still a lack of substantial data and standardized methodologies for TRUS images. In this thesis section, an approach that can be employed in the absence of training data is proposed; the underlying concept mainly relies on the theory of deformable superellipses. With the only requirement of moderate user interaction, the developed methodology reliably segments the prostate from TRUS images.

To show the effectiveness of the overall workflow, as well as the feasibility of implementation in a real-world clinical scenario, an image fusion procedure which relies on image registration between TRUS and MRI was developed. Hence, a semiautonomous segmentation framework for prostate cancer from TRUS images has been successfully realized, without relying on a large-scale dataset. Furthermore, the proposed framework can be employed as an annotation tool to ease and speed up the construction of prostate segmentation datasets, easing the future development of fully automated methods. Finally, another direction to investigate comprehends deformable registration techniques to further improve the image fusion step.

3.4 Lung COVID-19 Characterization

The study described in this section aims to propose a method for characterizing the lung parenchyma and lesions of COVID-19 patients with respect to the onset of PTE. Accurate segmentation of these ROIs is the prerequisite for accomplishing this task. Contrarily to healthy lung parenchyma segmentation, which can be obtained with thresholding, the segmentation of the lesions caused by interstitial pneumonia is not an easy task, since the lesions may be confused with surrounding tissues and vessels. Both a healthy lung parenchyma model and a lesion model have been trained in order to correctly segment all of the lungs belonging to COVID-19 patients.

The first part of this study posed the basis for further analysis of the segmented region, which may be used to establish a relationship between lesions and the onset of PTE. In the second part of this research work, different slices were selected from each CT scan, and feature extraction has been performed in these regions. These radiomic features were used for characterizing lungs in PTE and non-PTE patients.

According to the state-of-the-art presented in Section 2.2.7, no other author has considered the problem of performing statistical analysis of radiomic features from lung and lesion regions with respect to the onset of PTE in COVID-19 patients.

3.4.1 Materials

The study considers 20 CT scans of patients diagnosed with COVID-19 from the Polyclinic of Bari. Nine of these patients developed PTE. CT acquisition parameters were the following: slice thickness 0.6 mm, tube voltage 120 kVp, rotation time 0.33 s, pitch 1.2, and acquisition time 2.94 s. Images were reconstructed with a slice thickness of 1 mm in mediastinal and lung settings. Expert radiologists provided COVID-19 lung lesions (8 CT scans) and lung parenchyma (4 CT scans) annotations.

Publicly available datasets have also been considered for this work. Jun et al. provided a collection of 20 CT scans from COVID-19 patients with annotations of the left lung, right lung, and lesions [389]. Two datasets are available at MedSeg (<http://medicalsegmentation.com/covid19/>, *last accessed: 9 October 2021*). The first consists of 100 axial CT slices, coming from openly accessible JPG images of more than 40 patients, whereas the latter is composed of 9 volumetric CTs taken from Radiopaedia (<https://radiopaedia.org/>, *last accessed: 9 October 2021*). It should kindly be noted that Radiopaedia images have already been considered in the dataset by Jun et al. [389].

Zaffino et al. described a procedure for performing segmentation of CT images in COVID-19 patients by employing unsupervised learning techniques [390]. The authors realized a publicly available dataset composed of 50 CT scans with related annotations for healthy lungs, ground-glass opacity (GGO), lung consolidation (LC), and denser tissues.

Lastly, a dataset of patients with other lung ailments, VESSEL12 (<https://vessel12.grand-challenge.org/>, *last accessed: 9 October 2021*), has been considered. It comes from a challenge organized in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI 2012), held in Barcelona, Spain, from 2 to 5 May, 2012.

The summary of considered datasets is reported in Table 3.15, whereas example images are illustrated in Figure 3.26. The materials herein presented have been exploited for the analysis described in Section 3.4.2.

3.4.2 Workflow

In order to realize the segmentation of lung parenchyma and lesions, two CNN models based on the V-Net architecture have been trained, but considering the 2.5D and 2D variants with the same Dice loss formulation provided by Altini et al. [56]. For the task of lesion segmentation, two classes were considered: GGO and LC, as also detailed in Section 3.4.1. For that case, a multi-Dice loss formulation was provided, obtained by averaging the Dice loss across the two classes. Binary classification soft Dice coefficient is defined as in Equation (3.32).

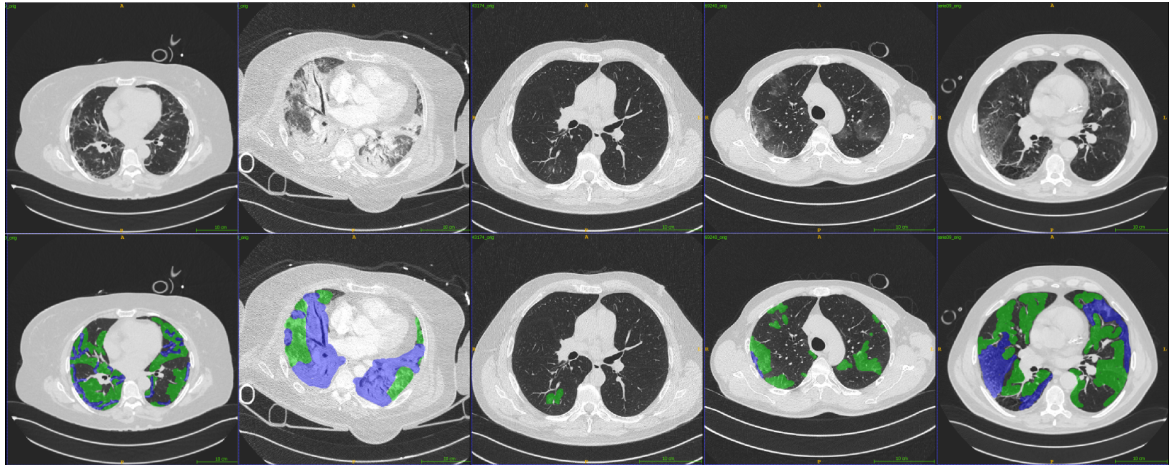


Fig. 3.26 *CT scans dataset of COVID-19 patients*. Top row: CT scans from the local dataset from the Polyclinic of Bari. Bottom row: annotations provided by expert radiologists. Green denotes ground-glass opacity (GGO), whereas blue indicates lung consolidation (LC).

Table 3.15 *COVID-19 CT scans datasets*. Datasets considered for this section. A stands for air, LP stands for lung parenchyma, LL stands for lung lesions, LC stands for lung consolidation, GGO stands for ground-glass opacity, and DT represents dense tissues.

Dataset	Acronym	Sample	Disease	Annotation
Jun et al. [389]	D1	20 CT	COVID-19	LP, LL
VESSEL12	D2	20 CT	Other lung diseases	LP
Ours	D3	20 CT	COVID-19	LP, GGO, LC
Zaffino et al. [390]	D4	50 CT	COVID-19	A, LP, GGO, LC, DT

$$D_S = \frac{\sum_{i=1}^N p_i g_i + \epsilon}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i + \epsilon} \quad (3.32)$$

Then, binary classification Dice loss can be defined as in Equation (3.33).

$$DL_S = 1 - D_S \quad (3.33)$$

Multi-class classification Dice loss can be defined as in Equation (3.34):

$$MDL_S = \frac{1}{K-1} \sum_{j=1}^{K-1} DL_{S_j} \quad (3.34)$$

where N represents the number of voxels in the batch, ϵ shows a constant for avoiding numeric errors, and K is the number of classes. It should be noted that the background class is not considered in the MDL_S calculation ($j = 0$ is not in the sum). In order to train a model capable of segmenting both lung parenchyma and lesions, the workflow followed is as depicted in Figure 3.27. For the lung segmentation model, images were pre-processed by considering a HU values range of $[-1250, 250]$, clamping values out of this range, and then rescaling into $[0, 1]$ for easing the training procedure.

A train set of 32 CT scans has been realized from the three datasets D1, D2 and D3 introduced in Table 3.15. In particular, 8 CT scans have been randomly selected from D1, 16 CT scans from D2, and 8 CT scans from the local dataset D3. The training sets are denoted as T1, T2, and T3, respectively. As a validation set, 2 CT scans have been randomly selected from D1, 4 CT scans from D2, and 4 CT scans from D3. These validation sets are denoted as V1, V2, and V3, for an overall validation set of 10 CT scans. A lung segmentation model was trained with 2.5D V-Net on the T1 and T2 train sets, whilst a lung lesion segmentation model was trained with 2D V-Net on the T3 train set.

The models were validated in a 'combinate' usage on the final validation set. Namely, the masks obtained from both models were merged, exploiting the logical OR operation, resulting in a final lung mask. Lastly, a post-processing stage composed of the following three operations was performed. First, a binary morphological opening was carried out on the predicted mask by adopting a ball kernel with a radius of 5 voxels in order to remove noisy points. Afterward, only the two largest connected components were retained, considering that two lungs appear in the human body. Finally, a binary morphological closing operation was carried out to fill holes in the image using a ball kernel with a radius of 10 voxels.

3.4.3 Radiomic Features

Radiomics refers to the high-throughput extraction of several features from radiographic images, and it is more commonly used for characterizing solid cancers [8]. The important applications include non-small-cell lung cancer [391], head-and-neck cancer [14], glioblastoma [392], hepatocellular carcinoma [173], and breast cancer [393] among many others.

To implement the feature extraction pipeline, the PyRadiomics library has been exploited [318], which defines most of the image features in compliance with those prescribed by the Imaging Biomarker Standardization Initiative (IBSI) [319]. Eventual differences are explicitly stated in the PyRadiomics documentation.

From a radiological point of view, any fundamental CT alteration (ground-glass opacity, consolidation, linear opacity, nodules) can be found in COVID-19 interstitial pneumonia as well as in neoplastic lesions. Tumors are usually discrete focal lesions, while in inflammatory conditions, such as COVID-19 pneumonia, the same findings may be widely and bilaterally distributed. Therefore, the different distribution of the lesions is decisive for a correct interpretation.

According to Gillies et al., radiomic features can be distinguished into two categories: “semantic” and “agnostic” features [394]. The first refers to those that are generally adopted by radiologists to describe ROIs, whereas the latter are those devoted to capturing lesion heterogeneity via quantitative descriptors. Throughout this thesis, agnostic features have been considered. These features include first-, second-, or higher-order statistical indicators.

First-order statistical indicators refer to histogram-based features, which do not take into account the spatial relationships of voxels but only consider the distribution of individual voxels. Starting from an ROI, it is possible to extract from the histogram descriptors, such as mean, median, energy, entropy, minimum, maximum, percentiles, interquartile range, skewness (an indicator of asymmetry), kurtosis (a descriptor of flatness), and so on.

Second-order statistical indicators are those usually referred to as “texture” features; they have the potential to explain spatial interrelationships between voxel intensity values. They can be derived for instance by the GLCM, which allows measuring the existence of voxels with the same intensities along a given direction, or from the GLRLM, which allows the quantification of repeated voxels with the same intensity along given directions. Starting from the GLCM, it is possible to extract features like correlation, contrast, auto-correlation, dissimilarity, energy, cluster tendency and shade, difference entropy, and so on. Instead, from the GLRLM, it is possible to extract features that describe short-run and long-run run emphasis, run length and gray-level non-uniformity, low and high gray-level run emphasis, run percentage, and so on [13].

Higher-order statistical indicators can be attained by applying statistical methods after having performed filtering or mathematical transformations to images. Instances of such transformations are wavelet transforms or Laplacian transform of Gaussian-filtered images, which can permit the extraction of zones with texture patterns that become more coarse [12].

Lastly, shape features are those that describe the geometric properties of the delineated ROI, such as maximum surface, volume, maximum diameter along diverse orthogonal directions, sphericity, and compactness. Shape features can be extracted from both 2D and 3D ROIs. It is worth noting that this kind of analysis can lead to the generation of feature vectors with hundreds of elements, with the risk of retaining redundant information. Moreover, the number of extracted features can be higher than the sample size, reducing the power and the generalization capability of the study. Therefore, the exploitation of feature selection or dimensionality reduction techniques is strongly recommended in order to generate highly informative, reproducible, and non-redundant feature vectors [13].

3.4.4 Results

3.4.4.1 Segmentation Experimental Results

From the experiments conducted, it was possible to obtain the metrics reported in Table 3.16. Example CT scans, ground truth, and predictions can be seen from Figure 3.28.

Table 3.16 *Experimental results for lung segmentation in COVID-19 patients from CT scans.* **LM** stands for lung model, **LLM** stands for lung model plus lesion model. **LLMPP** stands for lung model plus lesion model plus post-processing.

Model	Precision [%]	Recall [%]	Dice [%]	RVD [%]	ASSD [mm]	MSSD [mm]
LM	98.58 ± 0.92	95.58 ± 4.78	96.98 ± 2.46	-3.03 ± 5.16	1.68 ± 1.06	91.29 ± 33.67
LLM	98.35 ± 0.95	96.89 ± 3.42	97.58 ± 1.76	-1.47 ± 3.79	1.71 ± 0.92	96.21 ± 29.38
LLMPP	99.33 ± 0.49	95.77 ± 3.73	97.48 ± 1.98	-3.58 ± 3.86	0.95 ± 0.68	37.04 ± 27.03

It can be clearly observed that the lung model (**LM**) appears to have some trouble in predicting the whole lung region when it is highly lesioned. In order to overcome these difficulties, it was combined with a lesion model, resulting in the lung model plus lesion model (**LLM**) that is more likely to individuate the lung correctly with lesioned regions. Since the human body carries only two lungs, and they form a closed structure from a morphological point of view, applying post-processing based on morphological operators and connected component analysis can offer several benefits. In fact, it can be seen that

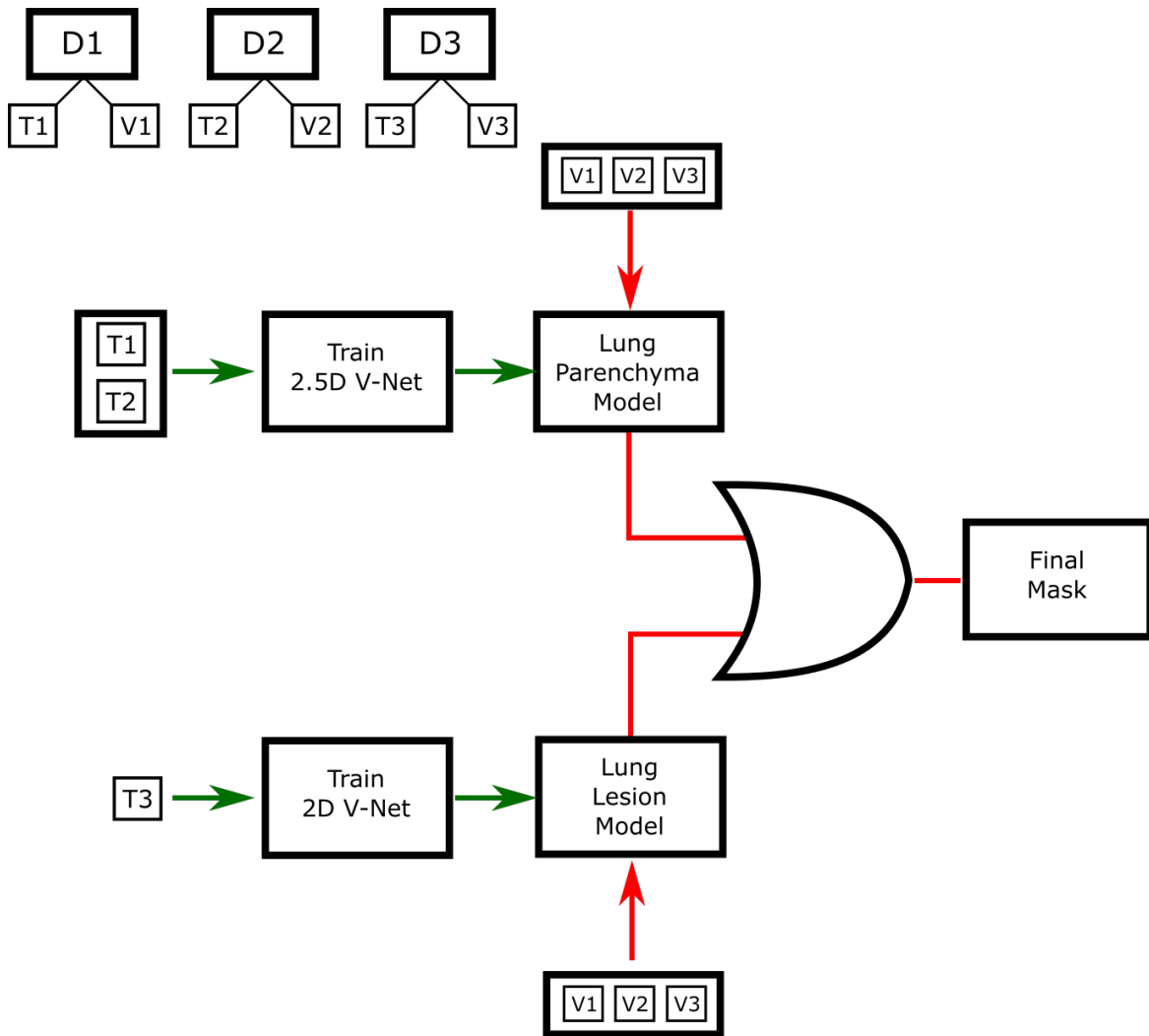


Fig. 3.27 Workflow for segmentation of lung parenchyma and lesions. CT scans, coming from three datasets, are clipped to $[-1250, 250]$ and then normalized by rescaling in the range $[0, 1]$. Two separate models are trained, a 2.5D V-Net exploiting T1 and T2, and a 2D V-Net exploiting T3. Final results are obtained by merging the masks from the two models with the logical OR operation. Metrics are assessed on the union of V1, V2, and V3 validation sets.

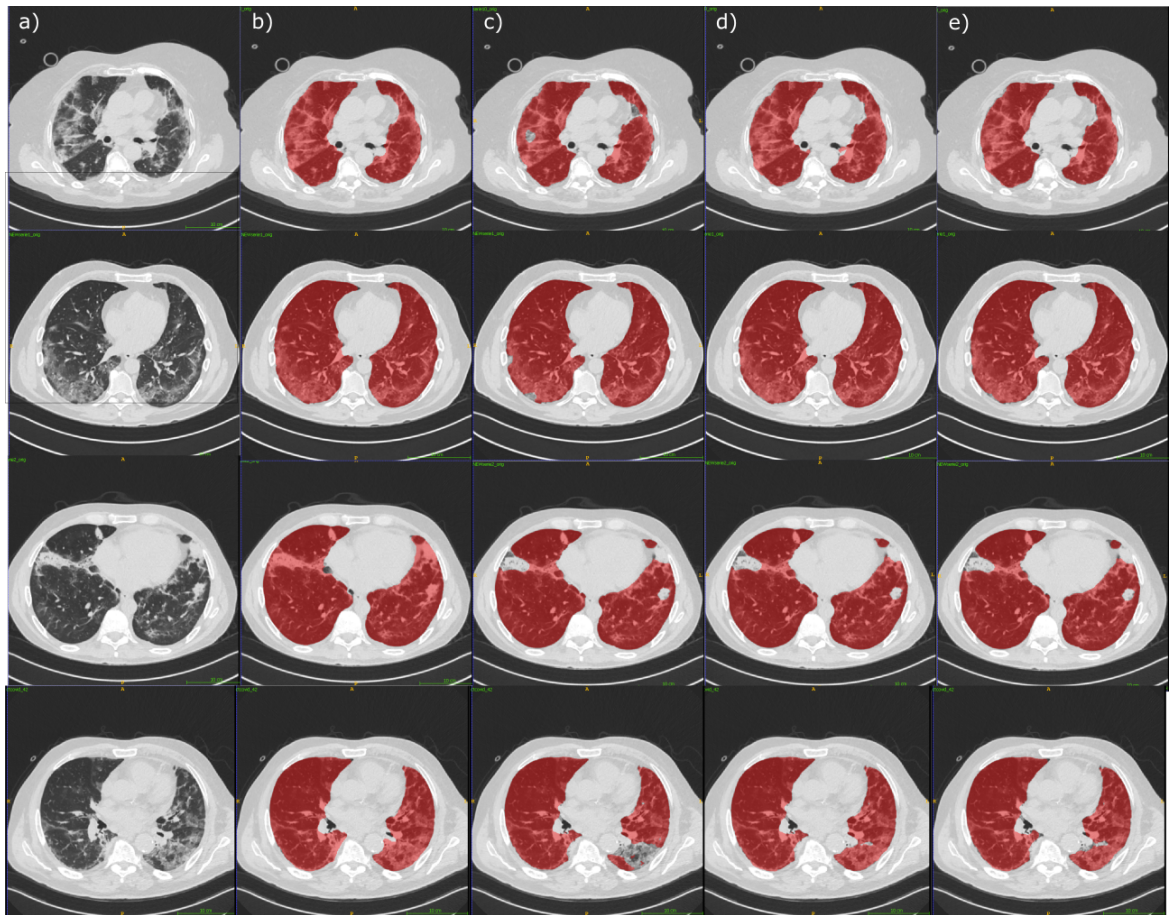


Fig. 3.28 *Example CT scans and predictions.* (a) original CT scans; (b) ground truth from radiologists at the Polyclinic of Bari; (c) lung model predictions; (d) lung model predictions combined with lesions predictions; (e) lung model predictions combined with lesions predictions, after post-processing.

MSSD is reduced from an average of more than 90 mm to an average of 37 mm with the lung model plus lesion model plus post-processing (**LLMPP**).

3.4.4.2 Radiomics Lung Characterization

In this phase, the features extracted from the 20 slices containing the largest portions of the lung from each CT scan have been investigated in order to detect differences in feature distribution between patients who developed PTE or not. In order to obtain the mask of the lung, the methodology for semantic segmentation developed in Section 3.4.2 of this thesis has been exploited.

According to the analysis of the state-of-the-art, the possibility to characterize lung parenchyma and lesions with radiomics methodologies from COVID-19 patients, with respect to the onset of PTE, has never been explored before; therefore, this can be considered one of the main contributions of the current section of the thesis.

PyRadiomics, introduced in Section 3.4.3, has been configured to extract features by considering the original image, five versions of the image after having applied Laplacian of Gaussian (LoG), with sigma ranging from 1.0 to 5.0, and Wavelet versions of the image. Later, the subsequent classes of features were considered: 2D shape descriptors (*shape2D*), first-order statistics (*firstorder*), texture classes gray-level co-occurrence matrix (*glcm*) [20, 332, 333], gray-level run length matrix (*glrlm*) [395, 396], gray-level size zone matrix (*glszm*) [397], gray-level dependence matrix (*glzm*) [398], neighboring gray-tone difference matrix (*ngtgm*) [399].

CT scans were considered in a 2D slice-wise fashion on the axial plane, resampling x and y axes to a resolution of 1 mm, exploiting the *sitkBSpline* interpolator. According to PyRadiomics documentation, CT intensity values reflect absolute world values (HU), which should be comparable across scanners. In studies where multiple scanners are considered, researchers have to check if extracted features are correlated to the scanner adopted, to avoid the risk of models capturing batch effects. However, given that the considered cohort comes from the same hospital, these issues are not a reason for concern. Image discretization was performed with a bin width set to 25, often considered as default in PyRadiomics usage. This process resulted in the extraction of 919 features for 400 samples (20 slices for each of the 20 CT scans).

Features have been visualized in a low-dimensionality space, by both considering principal component analysis (PCA) [400] and t-SNE [401]. For further visual understanding, the result of the 2D embedding space with PCA and t-SNE space is portrayed in Figure 3.29. No clear pattern emerged from either PCA or t-SNE embedding plots starting from the

original feature space, but the feature representation improves after having selected a subset of uncorrelated and significant features.

Likewise, a feature reduction procedure has been performed by considering statistical Mann–Whitney U tests on the designated radiomic feature set. In particular, the features for which the corresponding corrected p -value was lesser than 0.01 (their distribution was tested for both PTE and non-PTE samples) have been retained, resulting in an array of 109 features. The Holm–Šidák correction was exploited to counteract the multiple comparisons problem [402].

Starting from these 109 features, the correlation matrix has been investigated with a cluster map, in order to assess the correlations between the features and eventually reduce the feature set. Cluster map performs hierarchical clustering. In the case under analysis, the unweighted pair group method with arithmetic mean (UPGMA) [403] clustering algorithm was used, with Euclidean distance as metric. To visualize the cluster map, kindly refer to Figure 3.30. Indeed, it is possible to see that there are clusters of correlated features. In particular, many features of the groups *glcm*, *glrlm*, *glszm* show high intra-family correlation. Consequently, features with a correlation higher than 0.5 were removed, resulting in nine uncorrelated radiomic features that can be considered for the realization of a prognostic signature related to PTE assessment in COVID-19 patients. The procedure exploited for removing correlated features is reported in Algorithm 1.

The nine radiomic features proposed to compose a prognostic signature are:

1. *original_glcm_MaximumProbability*, defined in Equation (2.29);
2. *original_glszm_GrayLevelVariance*, defined in Equation (2.38);
3. *original_gldm_DependenceEntropy*, defined in Equation (2.41);
4. *log-sigma-1-0-mm-3D_glcm_InverseVariance*, defined in Equation (2.31);
5. *log-sigma-2-0-mm-3D_glrlm_GrayLevelVariance*, defined in Equation (2.43);
6. *log-sigma-2-0-mm-3D_glrlm_LongRunLowGrayLevelEmphasis*, defined in Equation (2.45);
7. *log-sigma-2-0-mm-3D_ngtdm_Contrast*, defined in Equation (2.47);
8. *log-sigma-4-0-mm-3D_firstorder_Skewness*, defined in Equation (2.48);
9. *wavelet-HH_glcm_ClusterShade*, defined in Equation (2.32).

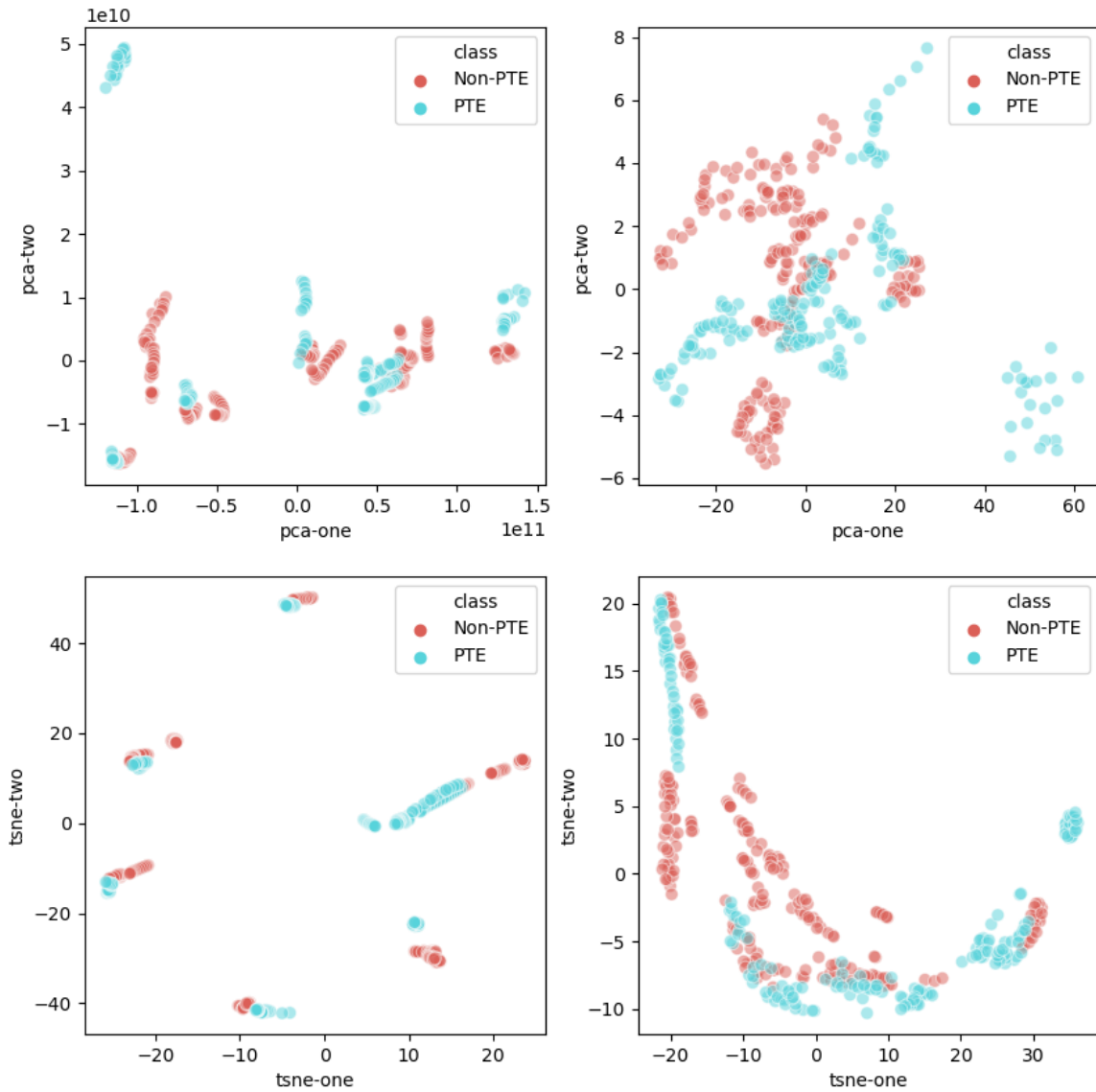


Fig. 3.29 Radiomic features in the 2D PCA and t-SNE spaces. (Top left) PCA starting from the 919 features; (top right) PCA starting from the radiomics signature of nine uncorrelated and significant features; (bottom left) t-SNE starting from the 919 features; (bottom right) t-SNE starting from the radiomics signature of nine uncorrelated and significant features.

Algorithm 1: Correlated Features Removal.

```

input :  $X$ , the original dataset, an  $N \times P$  matrix
          $\varepsilon$ , the correlation threshold // set to 0.5
output :  $LC$ , the list of correlated features to remove

 $LC = \{\}$ 
 $R_{XX} = \text{corr}(X, X)$  //  $R_{XX}$  is a  $P \times P$  matrix

 $i = 0$ 
while  $i < P$  do
     $j = 0$ 
    while  $j < i$  do
        if  $(R_{XX}[i, j] \geq \varepsilon) \wedge (j \notin LC)$  then
             $LC = LC \cup \{i\}$ 
        end
         $j = j + 1$ 
    end
     $i = i + 1$ 
end

```

These feature names, derived from PyRadiomics, are composed of three parts, divided by the underscores.

The first part refers to the image from which the features were extracted. It could be the original image (*original*), the image after having applied LoG with a given sigma (*log-sigma*-), or the image after having applied wavelet transform (*wavelet*-). In particular, four decompositions per level were obtained with Coiflets [404, 405] from 2D slices; that is, all possible combinations of applying either a high or a low pass filter in each of the two dimensions. The parameter N of the Coiflets was set to 1.

The second part refers to the kind of features, as first-order statistical indicators (*firstorder*), second-order statistical indicators (as *glcm* or *glrlm*), and so on.

The last part refers to the specific feature name inside its family, as defined by the PyRadiomics documentation [318], which can be consulted for further information. Appendix 2.2.7.1 summarizes mathematical details and definitions for the features exploited in this study.

Violin plots showing the difference between distributions for PTE and non-PTE cases are reported in Figure 3.31. Statistics concerning the distributions of these features for PTE and non-PTE cases are reported in Table 3.17.

Table 3.17 *Statistics for feature distribution between PTE and non-PTE cases.* Std is the standard deviation and IQR is the interquartile range. U_{PTE} and $U_{Non-PTE}$ are the two-sided U statistics.

Feature	Mean		Median		Std		IQR		U_{PTE}	$U_{Non-PTE}$
	Non-PTE	PTE	Non-PTE	PTE	Non-PTE	PTE	Non-PTE	PTE		
1	0.0485	0.0683	0.0390	0.0519	0.0322	0.0430	0.0325	0.0644	24999	14601
2	749.736	843.911	742.459	821.026	119.081	218.129	110.752	218.576	25849	13751
3	68.824	66.891	71.056	66.667	0.5538	0.4505	0.6482	0.5939	14637	24963
4	0.4233	0.4029	0.4162	0.4032	0.0210	0.0258	0.0269	0.0389	11459	28141
5	142.651	137.572	148.364	132.101	26.378	20.881	30.981	22.403	14731	24869
6	0.0137	0.0164	0.0123	0.0160	0.0053	0.0057	0.0079	0.0065	25470	14130
7	0.0088	0.0078	0.0087	0.0074	0.0021	0.0024	0.0027	0.0041	14680	24920
8	-0.3484	-0.1261	-0.2963	-0.1591	0.2789	0.3489	0.3635	0.4178	25838	13762
9	0.6364	-0.4777	-0.0010	-0.0092	15.061	24.794	0.0544	0.0322	14646	24954

Table 3.18 *Literature overview for COVID-19 image tasks.* Considered tasks include segmentation, detection, and classification. CT stands for computed tomography, ROC stands for receiver operating characteristic, IoU stands for intersection over union, AUC stands for area under the curve, AP stands for average precision, NPV stands for negative predictive value, PR stands for precision-recall, MCC stands for Matthews correlation coefficient, CNN stands for convolutional neural network, FCN stands for fully convolutional network, R-CNN stands for region-based convolutional neural network, DPN stands for dual path network, FRAGL stands for fuzzy region-based active contours driven by weighting global and local fitting energy.

Author	Method	Materials	Task	Results
Wu et al. [307]	Joint classification and segmentation (JCS)	144,167 CT images (3,855 annotated)	Real-time and explainable COVID-19 classification	Classification: sensitivity = 0.95, specificity = 0.93; segmentation: Dice = 0.78
Akbari et al. [308]	Active Contour	100 CT images	COVID-19 lesions segmentation	FRAGL: Dice = 0.96, Jaccard = 0.93, F1 = 0.66, precision = 0.91, recall = 0.53
Cao et al. [309]	U-Net-like CNN	2 CT scans	Objective assessment of pulmonary involvement and therapy response in COVID-19	Qualitative
Rajinikanth et al. [310]	Firefly algorithm and multi-thresholding based on Shannon entropy + Markov Random Field segmentation	100 CT images	COVID-19 lesions segmentation	Jaccard = 0.84, Dice = 0.89, accuracy = 0.92, precision = 0.92, sensitivity = 0.95, specificity = 0.94, NPV = 0.93
Rajinikanth et al. [311]	Harmony search optimization and Otsu thresholding	90 CT coronal slices + 20 CT axial slices	COVID-19 lesion segmentation	Infection rate
Ter-Sarkisov et al. [312]	One shot model based on Mask R-CNN	750 CT images for segmentation + 1492 for classification	COVID-19 classification and lesions segmentation	Segmentation: AP@0.5 IoU = 0.614, Classification: accuracy = 0.91
Zhao et al. [313]	Multi-task learning and self-supervised learning	349 COVID-19 CT images + 463 non-COVID-19 CT images	COVID-19 classification	F1 = 0.90, AUC = 0.98, accuracy = 0.89
Wang et al. [314]	Segmentation: FCN, U-Net, V-Net, 3D U-Net++; Classification: ResNet-50, inception, DPN-92, attention ResNet-50	1,136 CT images	COVID-19 pneumonia detection	Sensitivity = 0.97, specificity = 0.92, AUC = 0.991
Oulefki et al. [315]	Improved Kapur entropy-based multilevel thresholding procedure	275 CT images from COVID-CT-dataset + 22 CT images (local)	COVID-19 lesions segmentation	Accuracy = 0.98, sensitivity = 0.73, F1 = 0.71, precision = 0.73, MCC = 0.71, Dice = 0.71, Jaccard = 0.57, specificity = 0.99
Zheng, Wang et al. [316, 317]	U-Net + 3D deep neural network	499 CT volumes (train) + 131 CT volumes (test)	Lung segmentation + COVID-19 classification	ROC AUC = 0.96, PR AUC = 0.98, sensitivity = 0.91, specificity = 0.91

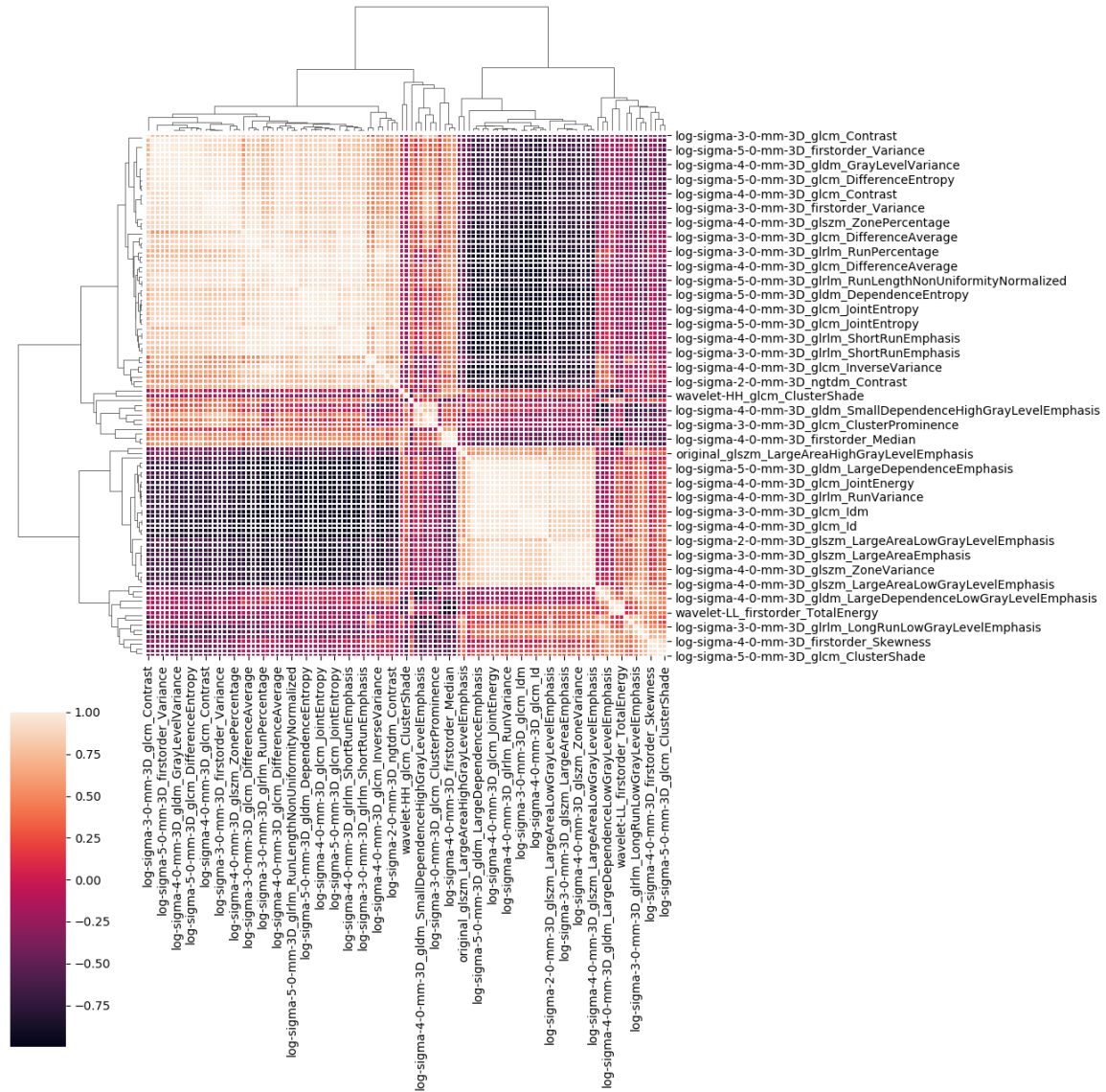


Fig. 3.30 Cluster map for the radiomics features. 109 radiomic features have been considered, with corrected p -value lesser than 0.01.

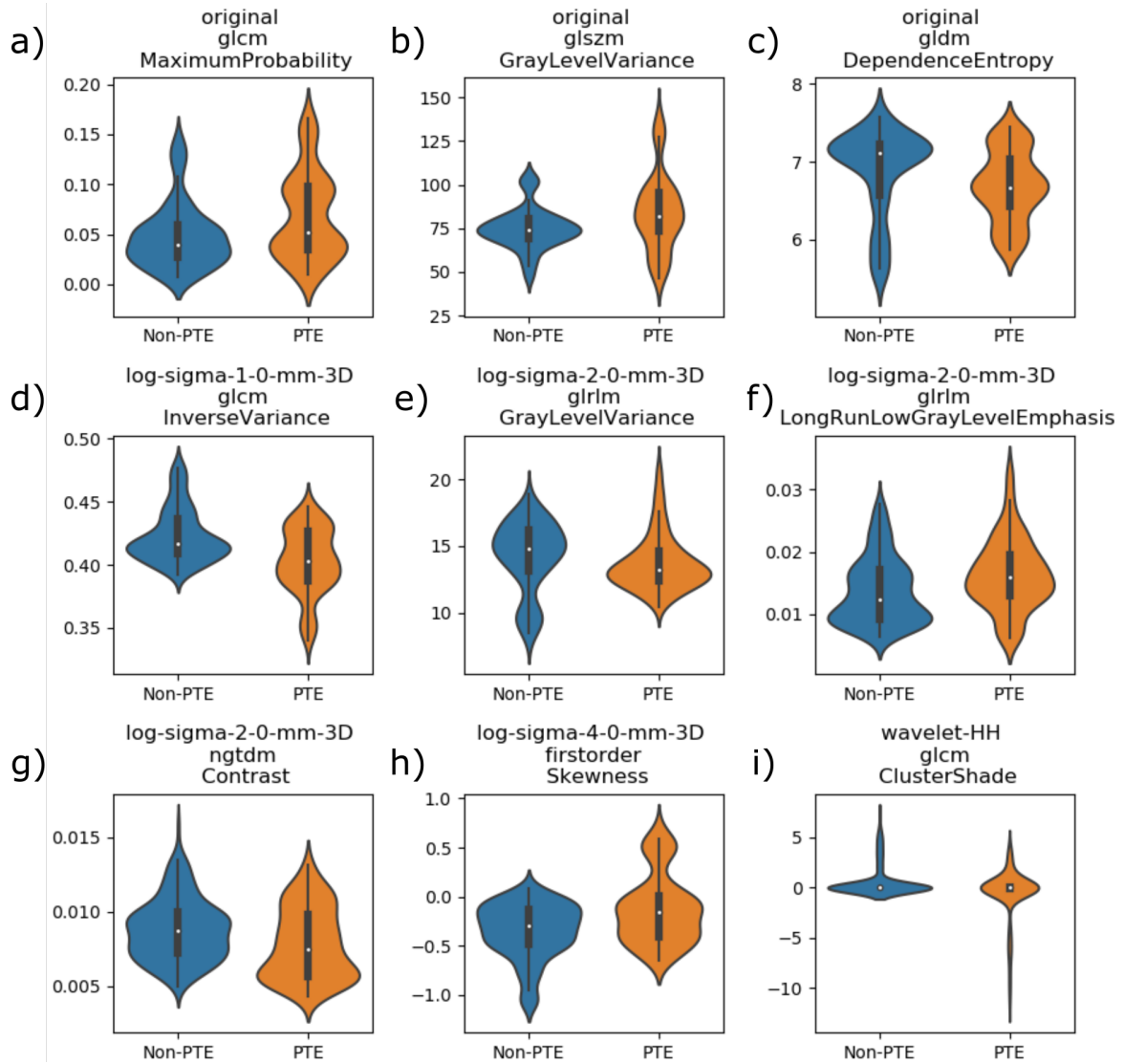


Fig. 3.31 Violin plots for the subset of 9 significant and uncorrelated features. (a) *glcm* maximum probability of the original image; (b) *glcm* gray level variance of the original image; (c) *gldm* dependence entropy of the original image; (d) *glcm* inverse variance of the LoG (Laplacian of a Gaussian) version of the image with $\sigma = 1$; (e) *glrlm* gray level variance of the LoG version of the image with $\sigma = 2$; (f) *glrlm* long run low gray level emphasis of the LoG version of the image with $\sigma = 2$; (g) *ngtdm* contrast of the LoG version of the image with $\sigma = 2$; (h) histogram skewness of the LoG version of the image with $\sigma = 4$; (i) *glcm* cluster shade of the HH wavelet version of the image. Scales of values on y axis have to be multiplied by a factor of 10 for the plots: (b), (c), (e).

3.4.4.3 Discussion

Several authors have already dealt with the problem of COVID-19 detection from CT scans [307, 312, 313, 316, 317] or focused on the task of lung and lesion segmentation [308, 310–312, 315–317].

The patients from the local cohort considered for this study were all diagnosed with COVID-19 since the objective was not that of realizing another algorithm for COVID-19 detection. The methods, materials, and results adopted by the related works are summarized in Table 3.18. It is possible to see that COVID-19 classification algorithms manage to obtain high performances. In regard to the segmentation, it is worth noting that some authors did not provide quantitative comparable metrics [309, 311]. Other authors provided metrics for lesions' segmentation, but not for lung parenchyma and lesion joint segmentation. Nonetheless, it is possible to see from the performed experiments that the proposed model is better at segmenting lungs with respect to models that have been trained only on non-COVID-19 patients.

According to the state-of-the-art analysis, no other work in the literature considered the possibility of exploiting radiomics methodologies in order to characterize the lungs between PTE and non-PTE patients. As can be seen from Figure 3.31 and Table 3.17, the designated features show different characteristics, from a statistical point of view, among the two conditions. This can aid in posing the basis for the development of a reliable predictive model, which will need to be tested on external and larger validation sets. Instead, in this thesis work, the main focus was on the feature selection and statistical analysis stage.

3.4.5 Summary

The COVID-19 pandemic has led to drastic impacts on the world. To address the deadliest complications of COVID-19, CT scans are routinely adopted in diagnostic and prognostic procedures. The task of manually segmenting lung parenchyma and lesions is tedious and error-prone, and it can suffer from intra- and inter-operator variability. Therefore, an automated workflow for segmenting lung lesions and parenchyma has been proposed, aiding the work of radiologists and posing the basis for further analysis of the segmented region. The segmentation stage achieved accurate results with a mean Dice coefficient of 97.48% and a mean ASSD of 0.95 mm for the chosen model.

In this preliminary study, 20 CT scans have been collected from the Polyclinic of Bari, all from patients positive for COVID-19. Nine of these patients developed PTE, which is a very unfavorable complication of COVID-19. To construct the ground truth, lung parenchyma

and lesions have been labeled by expert radiologists for 4 and 8 CT scans, respectively. Combining these annotations with publicly available datasets, a training set of 32 CT scans, and a validation set of 10 CT scans have been built.

CT scans are the clinical routine procedure to assess the severity of lung lesions but do have not the same sensitivity as CTA scans to diagnose PTE. In this study, a radiomic signature has been extracted to characterize the patients with a higher risk of PTE onset. The outcome of this stage consisted of the derivation of nine uncorrelated features that show significantly different statistical distributions according to the Mann–Whitney U test. However, more data need to be acquired in order to better understand the generalizability of the discovered feature set, which is considered an interesting direction for future work.

Chapter 4

Pathomics and Applications of Deep Learning in Digital Pathology

4.1 Stain Color Normalization in CRC Histopathology

A reliable pipeline for stain color normalization and tissue classification in H&E samples of patients with CRC is presented in this section. For the stain color normalization stage, five GANs based on the UI2IT framework have been thoroughly compared. Furthermore, four traditional image processing normalization techniques [406–410] have been considered as baselines. In order to assess the feasibility of the proposed solution, an evaluation of the generated tiles has been realized by an expert pathologist, introducing a metric defined as Pathologist Perceptive Quality (PPQ). Afterward, three CNNs have been considered as feature extractors from tile normalized with the previously mentioned techniques. An SVM has been trained on top of deep features, in order to assess the classification accuracy of the developed CAD system.

Contrarily to what is usually done in stain color transfer, where a generative model is trained between each pair of domains, a meta-domain — TCGA — composed of the union of data coming from several laboratories, has been considered in place of the source domain in the training phase of the stain color normalization module. With the only need to eventually perform a double normalization at inference time, on both source and target classification domains, the proposed methodology has the advantage of avoiding the expensive process of training multiple GANs.

Summarizing, in the work proposed in this section, the following innovative contributions are introduced: (i) an extensive comparison of normalization techniques in order to assess the most reliable ones for validating tissue classifiers on data coming from different laboratories;

(ii) an investigation of features extracted from deep CNN architectures for CRC tissue classification; (iii) an evaluation method for assessing the quality of generated tiles from expert pathologists, resulting in the conceptualization of a novel perceptive metric, PPQ; (iv) a setup for UI2IT stain color normalization which does not require to train style transfer GANs between every pair of data domains; (v) a collection of a validation cohort of samples enrolled at the IRCCS Istituto Tumori Bari “Giovanni Paolo II”, resulting in a publicly available dataset of 10 WSIs and N=15,857 annotated tiles [411].

Section 2.3.1.1 details related works in both colorectal cancer tissue classification and stain color normalization. The remainder of this section is structured as follows. Section 4.1.1 presents materials and methods adopted for this research. Firstly, the datasets collected and employed for the study are described. Then, image processing and deep learning-based normalization techniques are presented, with a special focus on UI2IT settings. Evaluation metrics for assessing the quality of generated normalized images are described, with specific considerations for histopathological scenarios. The experimental UI2IT setting, which exploits a meta-source dataset, is introduced. Section 4.1.2 presents the experimental results, both for the quality of normalized images and for the classification task, which are then discussed in Section 4.1.3. Finally, conclusions and directions for future works are portrayed in Section 4.1.4.

4.1.1 Materials and Methods

4.1.1.1 Datasets

Two kinds of datasets have been collected for this research: datasets for stain color normalization and for multi-class classification. All datasets contain histopathological image tiles belonging to WSIs of patients diagnosed with CRC.

For what concerns stain color normalization, the considered datasets are: (SD1) a dataset introduced by Kather et al. [412]; (SD2) a local dataset collected at IRCCS Istituto Tumori “Giovanni Paolo II”. The SD1 dataset is composed of image tiles coming from 604 CRC WSIs in the TCGA database, while the SD2 dataset is composed of image tiles coming from 58 WSIs. All tiles have dimensions of 512×512 px at $0.5 \mu\text{m}/\text{px}$. SD1 and SD2 have been used to construct two tile-level datasets, each with a training set of N=100,000 tiles and a test set of N=50,000 tiles.

With respect to the multi-class classification, the considered datasets are: (CDT) a dataset introduced by Kather et al. [25, 412]; (CDV) a local dataset collected at IRCCS Istituto Tumori “Giovanni Paolo II”.

The CDT dataset [25, 412] is composed of $N=100,000$ image tiles from H&E histological tissue of humans with CRC, subdivided into nine tissue classes. The size of the images is 224×224 pixels, which corresponds to $112 \times 112 \mu m^2$.

The CDV dataset is composed of $N=15,857$ tiles coming from 10 WSIs. The tiles have dimensions of 224×224 pixels which correspond to $116 \times 116 \mu m^2$. The dataset has been annotated by an expert pathologist and has been made publicly available [411], to ease the development and comparison of computational techniques for CRC histological image analysis.

Both the datasets have been classified into the following seven classes, as done in a precedent work [20]: *TUM* – tumor epithelium; *MUSC_STROMA* – the union of SIMPLE_STROMA, encompassing smooth muscle, tumor stroma, and extra-tumor stroma, and COMPLEX_STROMA, consisting of stroma or smooth muscle containing single tumor cells and/or few, non-aggregated immune cells; *LYM* – lymphoid follicles and other immune-cell conglomerates; *DEBRIS_MUCUS* – hemorrhage, mucus, and necrosis; *NORM* – normal mucosa; *ADI* – adipose tissue; *BACK* – background.

To accomplish this categorization for the CDT dataset, the *DEBRIS_MUCUS* class has been built by combining the DEB and MUC classes; instead, the *MUSC_STROMA* class has been constructed by fusing the MUS and STR classes. The CDT dataset obtained with this procedure is composed of $N=77,805$ tiles, considering that only half of the images of the merged classes have been retained, for maintaining class balancing during training.

For both the locally collected datasets, namely, SD2 and the CDV, the Institutional Ethics Committee of the IRCCS Istituto Tumori “Giovanni Paolo II” approved the study (Prot n. 780/CE).

Figure 4.1 portrays frequency distribution and example tiles from normalization and classification datasets.

4.1.1.2 Image Processing Normalization

Histopathology involves a manual staining procedure for preparing tissues prior to microscopic imaging for diagnosis. This is a non-standardized procedure that may cause considerable variability in the color characteristics of tissue samples from different laboratories; this can occur due to inconsistent tissue staining, color responses to different scanners, or differences in raw materials and stain manufacturing techniques. Stain color variation degrades the performance of CAD systems. In the presence of severe color variation in histopathological images, stain color normalization, which is achieved by removing the stains for visual enhancement, is a common practice.

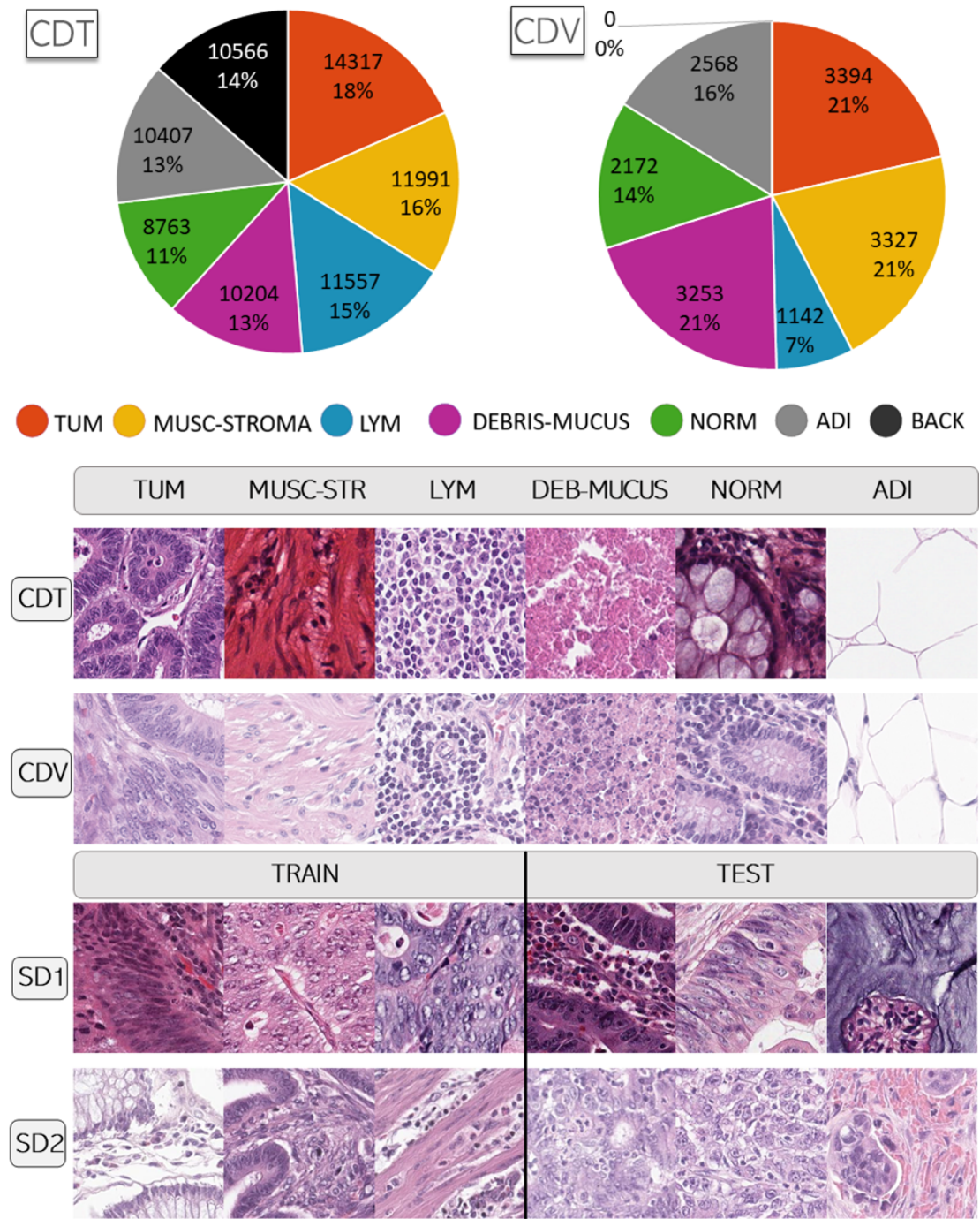


Fig. 4.1 *Datasets exploited for the stain normalization and the downstream classification task.* SD1 and SD2 are the datasets for stain normalization. CDT and CDV are the datasets for the classification task.

Among the most popular image processing methods, it is worth mentioning the works of Reinhard et al. [406], Macenko et al. [407], Khan et al. [408], and Vahadane et al. [409]. These works are usually considered as reference methods when authors propose novel methods for stain color normalization [326, 330, 336, 413, 414]. Given their importance as a benchmark, it has been decided to include all of them as baseline methods. Nevertheless, important limitations of these methods include the fact that they require H&E staining, the need of exploiting a template patch for fitting the stain distributions [21], and the introduction of color artifacts, as those portrayed in Figure 4.2.

For more details about normalization techniques and pre-processing procedures, the interested reader is referred to dedicated surveys [51, 415] or to the original papers mentioned before.

The proposed stain normalization methodology, based on UI2IT, can be applied to every type of staining and is not restricted to H&E images only.

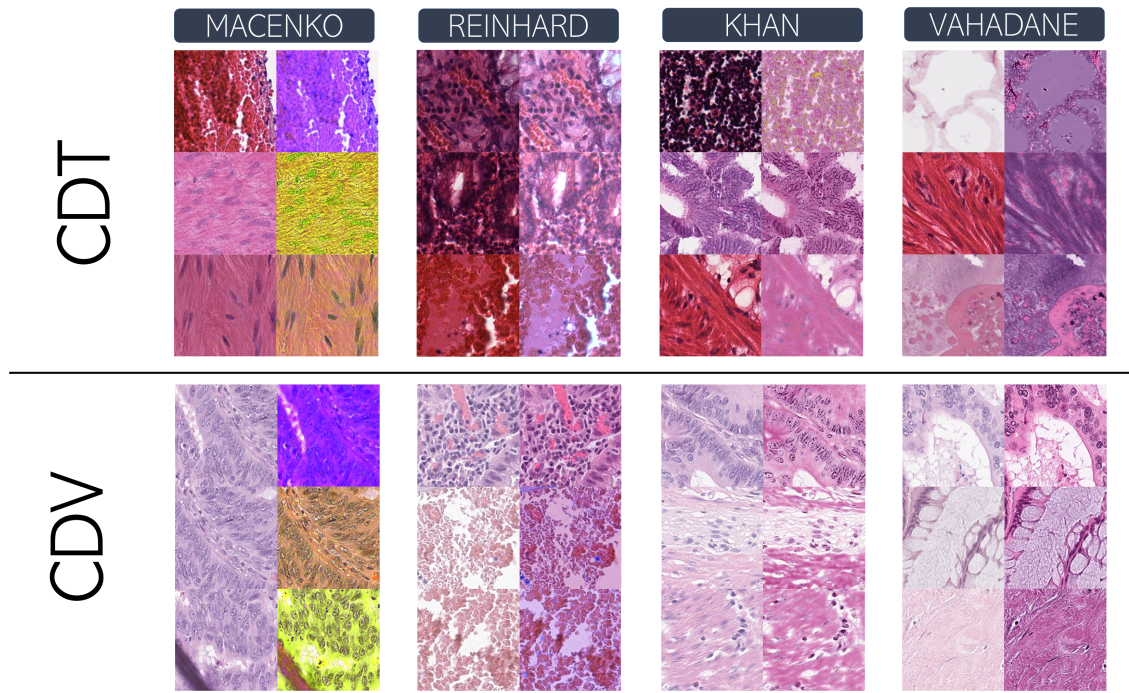


Fig. 4.2 Example of artifacts introduced by classical stain color normalization methods. (Top Row) Artifacts on images of the training set for classification (CDT). (Bottom Row) Artifacts on images of the test set for classification (CDV).

4.1.1.3 Deep Learning Normalization

Stain normalization of histopathological images performed with GANs is gaining much attention recently [326], with a particular focus on CycleGAN [99] and its variants [21, 200]. Nonetheless, many other algorithms belonging to the UI2IT framework have not been explored for the stain color transfer task.

In this subsection, GAN theory is briefly reviewed. Then, relevant works based on GANs for UI2IT are summarized and categorized, to explain the role of the proposed system in the normalization scenario for histopathological images. Relevant metrics which can be used for assessing the quality of generated images are also presented so that both mathematical and perceptual evaluations can be done. Also, the PPQ, a novel metric based on pathology assessment to estimate the quality of generated images in histopathological environments, has been introduced.

Interested readers may find useful information also in surveys concerning adversarial-learning-based I2IT [416–418] or GAN applications, techniques for training, and architectures [29, 100, 102, 419].

GAN and cGAN Goodfellow et al. introduced GANs in 2014 [419]. The GAN framework involves the simultaneous training of two models: a model which has to capture the underlying data distribution, the generator G , and a model, the discriminator D , that has the responsibility to determine the odds that a sample belongs to those generated by G rather than the training distribution. G is trained by maximizing the possibility of D to make a classification error. This results in a minimax game with two players, where an optimum can be obtained with the generator G that recalls data from the underlying training distribution and the discriminator D which always predicts $\frac{1}{2}$. Both models can be trained taking advantage of backpropagation when they are neural networks. A general introduction to the GAN architectures is presented in Section 2.1.4. Here, the general definitions are contextualized with respect to the stain color normalization task.

Isola et al. faced the paired image-to-image translation task with GANs [98]. The basic idea behind their framework is that many computer vision challenges are handled with different algorithms even though the underlying problem is the same: converting pixels to pixels. The architecture they introduced, pix2pix, employs a cGAN for addressing these kinds of tasks. This enables learning also the loss function to train this map, differently from the basic learning of a transformation from an input image to an output one. In pix2pix, the discriminator D , during the training process, is devoted in distinguishing between a true pair,

(s, t) , and a fake pair $(s, G_{ST}(s))$, instead of t and $G_{ST}(s)$, where S, T are two domains and (s, t) represents a pair with sensible elements for both the domains.

Pix2pix architecture, which only allows tackling paired I2IT tasks, has not been included in the experiments, since paired images across domains were not available, as happens with many real-world problems.

Cycle-Consistency Constraint In UI2IT setups, cycle-consistency is the most widely adopted method for imposing association [420]. This paradigm is grounded on the concept of retrieving also the reverse mapping from the target domain back to the source one. Furthermore, it enforces a check that a sample input image can be reconstructed. Among the most well-known architectures which fall into this category, it is worth mentioning CycleGAN [99], DualGAN [421], and DiscoGAN [422]. Due to its wide adoption in the stain color normalization domain [327–329, 413], only CycleGAN has been included in the experiments among the cycle-consistency-based networks.

The problem with the pix2pix framework, which also occurs in stain color normalization setups, is that it requires paired training data, which is often not available. CycleGAN overcomes these limitations by learning a mapping $G_{ST} : S \rightarrow T$ in a way that $G_{ST}(S)$ is indistinguishable from T exploiting an adversarial loss. To avoid the issues coming from the under-constrained mapping, at the same time, an inverse mapping $G_{TS} : T \rightarrow S$ is also learned. In this way, the cycle-consistency loss can enforce $G_{TS}(G_{ST}(S)) \approx S$ and $G_{ST}(G_{TS}(T)) \approx T$.

One-sided Translation Instead of enforcing cycle-consistency, it is possible to promote relationships belonging to the input to be similarly reflected in the output. For instance, patches that are perceptually similar inside an input image should retain their proximity in the output. TraVeLGAN [423], DistanceGAN [424], and GcGAN [425] allow one-way translation, so avoiding the need for a cycle-consistency. The problem is that they require relationships between full images, or with predefined distance functions. Among these networks, the GcGAN has been included in the experiments carried out for this section.

While the cycle-consistency framework needs to train two generators simultaneously, G_{ST} and G_{TS} , one-directed domain translation can be successfully achieved also by only preserving pairwise images' distances. An important limitation of both cycle-consistency and distance preservation is that they do not properly consider simple geometric transformations.

The idea of enforcing a geometry-consistency constraint in a UI2IT GAN framework comes from the work of Fu et al. [425]. According to the authors, considering a geometric transformation $f(\cdot)$, the images from the source domain should not be altered by the corre-

sponding generators G_{ST} and $G_{S'T'}$, where S' and T' are the domains obtained by applying $f(\cdot)$ to S and T , respectively.

From a mathematical perspective, considering a random sample s from original domain S , a proper geometric transformation $f(\cdot)$, and its inverse $f^{-1}(\cdot)$, the geometry-consistency constraint can be formulated as $f(G_{ST}(s)) \approx G_{S'T'}(f(s))$ and $f^{-1}(G_{S'T'}(f(s))) \approx G_{ST}(s)$. Since it is improbable that G_{ST} and $G_{S'T'}$ make errors in the same region, the generator models act as co-regulators for each other, thanks to the geometry-consistency constraint, thus improving over mistakes in local zones of their relative translations.

The 90 degrees clockwise rotation has been considered as the geometric transformation in the experiments performed for this study since it has been proven useful by the authors of the original paper.

Note that the one-sided stain color consistency paradigm is particularly interesting for color stain normalization, since the objective is usually to translate tiles from the original domain to the target one, and therefore two generators are not needed.

Patchwise Contrastive Learning Approaches CUT and FastCUT [420] have been proposed to solve the limitations encountered in relationship preservation-based architectures, by replacing cycle-consistency with the possibility of learning a cross-domain similarity function by maximizing mutual information between corresponding patches from images belonging to the source and target domains, without the need to depend on some predefined distance.

The architecture proposed by Park et al. [420] enforces positive (related) patches to map to nearby points in the learned feature space if compared to negative (unrelated) patches coming from the dataset. This framework allows one-sided translation in UI2IT setups, resulting in greater quality and less training time compared to cycle-consistency-like approaches.

Mathematically, it is possible to define: the query $v \in \mathbb{R}^K$, the positive $v^+ \in \mathbb{R}^K$, and N negative samples $v^- \in \mathbb{R}^{N \times K}$. In these definitions, v , v^+ , v_i^- are K -dimensional vectors, and v_i^- is the i -th vector from the matrix of negatives v^- . The cross-entropy loss can then be calculated, defining the probability of a positive example to be chosen over negatives, as reported in Equation (4.1).

$$l(v, v^+, v^-) = -\log \left(\frac{\exp\left(v \cdot \frac{v^+}{\tau}\right)}{\exp\left(v \cdot \frac{v^+}{\tau}\right) + \sum_{i=1}^n \exp\left(v \cdot \frac{v_i^-}{\tau}\right)} \right) \quad (4.1)$$

τ is a temperature defined as equal to 0.07 by the original authors. The goal of the CUT framework is to relate source and target image patches. In the considered context, the query

concerns the target, whereas the positive and negative samples concern corresponding and noncorresponding source patches, respectively.

In order to enforce this relationship, the authors selected J layers with whom they encoded the input images and passed them to a two-layers MLP H_j , creating a stack of features $\{z_j\}_J = \left\{H_j \left(G_{enc}^j(s)\right)\right\}_J$, where G_{enc}^j is the output of the j -th layer. In a similar way, they also encoded the output image in $\{\hat{z}_j\}_J = \left\{H_j \left(G_{enc}^j(G(s))\right)\right\}_J$. The objective is to match the corresponding source-target patches for each layer and use the other ones from the source image as the negatives. They defined the features of the positive patch as $z_j^q \in \mathbb{R}^{C_j}$ and the features of the negative patches as $z_j^{Q \setminus q} \in \mathbb{R}^{(Q_j-1) \times C_j}$, where $j \in \{1, 2, \dots, J\}$ indexes the selected layer and $q \in \{1, 2, \dots, Q_j\}$ indexes the spatial location. The number of spatial locations and the number of channels for layer j are referred to as Q_j and C_j , respectively.

The loss is calculated as reported in Equation (4.2).

$$L_{PatchNCE}(G, H, S) = \mathbb{E}_{s \sim S} \left(\sum_{j=1}^J \sum_{q=1}^{Q_j} l \left(\hat{z}_j^q, z_j^q, z_j^{Q \setminus q} \right) \right) \quad (4.2)$$

The final objective function can be defined as portrayed in Equation (4.3), where $L_{PatchNCE}(G, H, T)$ is the identity loss that enforces the generator to avoid unnecessary changes.

$$L_{GAN}(G, D, S, T) + \lambda_s L_{PatchNCE}(G, H, S) + \lambda_t L_{PatchNCE}(G, H, T) \quad (4.3)$$

The configuration with $\lambda_s = 1$ and $\lambda_t = 1$ corresponds to CUT, whereas the configuration $\lambda_s = 10$ and $\lambda_t = 0$ is referred to as FastCUT.

AI-FFPE [426] is a modification of CUT that adds some unique characteristics, like the presence of a Spatial Attention Block (SAB) in the architecture of the generator and a L_1 regularization factor in the objective function, that is calculated as shown in Equation (4.4), where $L_{reg}(G, S) = \|S - G(S)\|_1$.

$$L_{GAN}(G, D, S, T) + \lambda_{reg} L_{reg}(G, S) + \lambda_s L_{PatchNCE}(G, H, S) + \lambda_t L_{PatchNCE}(G, H, T) \quad (4.4)$$

The authors of AI-FFPE claim that in the frozen-to-FFPE (formalin-fixation and paraffin-embedding) translation task, their model outperforms generic image-translation networks. Particularly, they state that their modifications to the CUT architecture contribute to the artifact-correcting performance of their model.

4.1.1.4 Stain Normalization Workflow

In the standard pipelines developed for stain color normalization exploiting UI2IT frameworks, two domains are usually considered: a source domain S and a target domain T . The objective is to transform images of the source domain to the distribution of the target domain. Therefore, a generator, to learn a mapping from S to T , namely $G_{ST}(S) \approx T$, is the outcome of the training stage for stain color transfer.

Instead, in the framework proposed for the CAD for CRC classification, three domains have been considered: the meta-domain M , the source domain S , and the target domain T .

In this application, M is a composition of multiple subdomains $S_1, \dots, S_n$, covering a wide variety of stain color conditions, so that by learning $G_{MT}(M) \approx T$, the map $G_{MT}(S) \approx T$ can be approximated.

When referring to the application presented in this section, M is the domain of images coming from the TCGA, S is the domain of images coming from the training set for the classification, and T is the domain of images coming from the local cohort of IRCCS Istituto Tumori "Giovanni Paolo II". In all other contexts, S and T are the generic domains considered for I2IT.

The GAN models have been trained from scratch with a learning rate of 0.0002, for 5 epochs, with the Adam solver [256] and a batch size of 1. Input images were of size 256×256 . No other preprocessing was applied to the images.

4.1.1.5 Generated Image Tiles Quality Assessment

Quantitative evaluation of the quality of images generated by GANs is not an easy task, but different approaches have been proposed in the literature. Common quality objective measures for image similarity include PSNR and SSIM [427]. As noted by Zhang et al. [428], these metrics are shallow, simple functions, which cannot measure in a proper way the human perception. They proposed to realize a perceptual distance with deep features exploiting the VGG network, resulting in a Learned Perceptual Image Patch Similarity (LPIPS) metric. Unluckily, these measures can be exploited only if ground-truth images are available [416], so they are excellent candidates to assess results obtained with conditional GAN for paired I2IT tasks, but are not directly applicable in UI2IT settings, at least to assess the adherence of the mapped image to the target domain. On the other hand, these measures can be considered to evaluate the introduction of artifacts between original images and their translated version.

Scores obtained exploiting Inception-v3 network pre-trained on ImageNet, as Inception Score (IS) [429], and Fréchet Inception Distance (FID) [430] were introduced to overcome

this issue and allow an assessment of realness and heterogeneity of generated images, from the point of view of feature distribution. To determine the FID, two multivariate Gaussians are fitted on feature vectors obtained by embedding samples from the Inception network. Then, the Fréchet Distance [431] or the Wasserstein-2 distance [432] is calculated among these two gaussian distributions, as in Equation (4.5).

$$FID(r, f) = \|\mu_r - \mu_f\|_2^2 + \text{Tr} \left(\Sigma_r + \Sigma_f - 2(\Sigma_r \Sigma_f)^{\frac{1}{2}} \right) \quad (4.5)$$

where μ_r , μ_f represent the mean of the real and fake generated sample feature vectors, respectively, and Σ_r , Σ_f represent the covariance matrix of the real and fake generated sample feature vectors, respectively.

In the research community, there is not a wide agreement on how to evaluate unpaired image-to-image translation frameworks [433]. Therefore, the considered experimental design involves the adoption of several metrics among the considered datasets.

To assess the reliability of the fake images generated via GANs, the PPQ, a novel perceptive quality measure, has been introduced. In detail, an image tile and the related GAN-normalized versions of the same tile were shown to an expert pathologist, and for each normalization method, the pathologist had to decide a discrete score between 1 and 4, where 1 represents an excellent quality image and 4 a bad one. In the experimental design considered for this study, 200 image tiles were shown to the pathologist, for both the SD1 and SD2 datasets.

4.1.1.6 Experimental Design

In order to perform a comprehensive comparison and analysis of normalization methods and downstream classifiers, the experimental design of the study has been realized as detailed in this section. G_{MT} refers to the generator that transforms images from meta-domain to target domain, whereas G_{TM} refers to the generator that performs the inverse mapping. In the proposed framework, besides training purposes for cycle-consistency-based methods, only G_{MT} is required. For the classical normalization methods, a reference tile, R , has been used for template matching.

CDT and CDV refer to the training and test set for classification, respectively. N(CDT) and N(CDV) refer to their normalized version, that is $G_{MT}(CDT) = CDT_{MT}$ and $G_{MT}(CDV) = CDV_{MT}$ for the GAN normalization, and CDT_{RT} and CDV_{RT} for classical normalization, respectively. SD1 and SD2 refer to the stain normalization datasets. N(SD1) and N(SD2) refer to their normalized version, that is $G_{MT}(SD1) = SD1_{MT}$ and

$G_{MT}(SD2) = SD2_{MT}$ for the GAN normalization, and $SD1_{RT}$ and $SD2_{RT}$ for the classical normalization, respectively. Images from CDV and SD2 belong to the T domain, images from SD1 belong to the M domain, and images from CDT belong to the S domain.

Regarding what concerns the normalization procedure:

1. A reference tile has been used for the traditional normalization methods (Reinhard, Macenko, Khan, Vahadane). Both the CDT and CDV datasets' image tiles have been normalized when used for classification. In order to assess normalized image distributions for these methods, FID has been calculated between:

- (a) $N(SD1)$ and $N(SD2)$,
- (b) $N(CDT)$ and $N(CDV)$.

2. In order to obtain a model capable of mapping CDT and CDV image tiles to the same target domain, all the GANs have been trained by exploiting SD1 and SD2 training and test sets. To assess the quality of the generated image distributions, FID has been calculated between:

- (a) $N(SD1)$ and $SD2$,
- (b) $N(SD1)$ and $N(SD2)$,
- (c) $N(CDT)$ and CDV ,
- (d) $N(CDT)$ and $N(CDV)$.

These tests have been made in order to assess if a double normalization (that is, normalizing images by performing $G_{MT}(S)$ and $G_{MT}(T)$) would work better than a single normalization (i.e., normalizing images by performing $G_{MT}(S)$ only, while keeping T images unaltered), in the proposed setting.

3. In order to understand the quality of the normalized images with respect to the original ones, PSNR, SSIM, and LPIPS have been determined between:

- (a) $N(SD1)$ and $SD1$,
- (b) $N(SD2)$ and $SD2$.

4. In order to understand the quality of the GAN-generated images, from the pathologist perspective, PPQ has been calculated considering:

- (a) $N(SD1)$ and $SD1$,

- (b) N(SD2) and SD2.
- 5. In order to understand how UI2IT models affect the saturation of images, SSIM between saturation channels (in HSV color space) has been calculated between:
 - (a) N(SD1) and SD1,
 - (b) N(SD2) and SD2.

For what concerns the downstream classification, three CNNs have been exploited as feature extractors before training an SVM classifier. In detail, the employed SVM was a multi-class classification error-correcting output code (ECOC) model with a one-vs-one coding design and Radial Basis Function (RBF) kernel. The three CNNs considered are DenseNet201 [65], InceptionV3 [64], and VGG16 [41]. The SVM model has been trained on features extracted from N(CDT) and validated on features belonging to N(CDV).

A detail of the experimental design is pictorially represented in Figure 4.3.

4.1.2 Experimental Results

The results obtained from the experiments described in Section 4.1.1.6 are presented in this section. In order to better characterize the two components of the developed pipeline, two subsections are delineated. The first one deals with the results of the stain color normalization, whereas the second one presents the results of the multi-class classification.

4.1.2.1 Stain Color Normalization

A visual example of the different techniques based on UI2IT for stain color normalization is portrayed in Figure 4.4. It is possible to notice different color patterns for the various methods. In particular, FastCUT tends to generate tiles that have a higher saturation than the ones generated by other methods. This fact can be seen from Figure 4.5 and Figure 4.6. Boxplots in Figure 4.5 show that, for the SD1 dataset, N(SD1) images obtained with FastCUT are the ones with saturation more similar to the original version. This is because FastCUT is the method that performs the lesser desaturation. Instead, for the SD2 dataset, N(SD2) images obtained with FastCUT are the ones with the greatest dissimilarity to the original ones. This is because other methods do not saturate these images, instead FastCUT increases the saturation of input images.

Image similarity measures, namely PSNR, SSIM, and LPIPS, have been calculated between SD1 and N(SD1) and SD2 and N(SD2), in order to assess the presence of artifacts

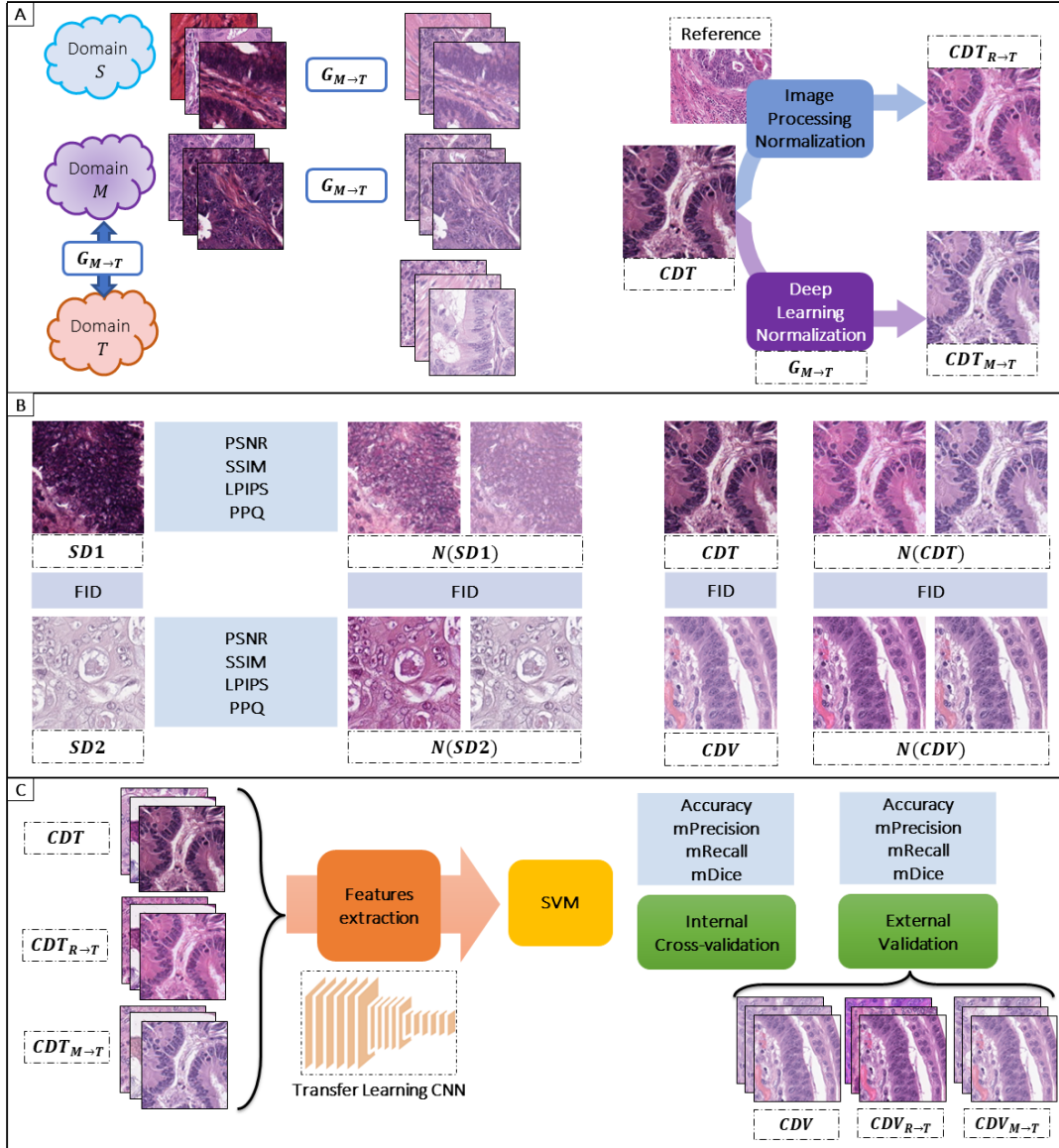


Fig. 4.3 Workflow employed for the stain color normalization study. (A) Training and use of the stain color normalization module. The GANs belonging to the UI2IT framework are trained by exploiting a meta-domain M . Source and Target domains, S and T , refer to the classification task. Traditional color normalization techniques exploit a reference tile R . (B) Assessment of the stain color normalization models. (C) Training and validation of the CRC tissue classifier.

or image degradation when performing stain color normalization. These quantitative results are reported in Table 4.1 and Figure 4.7.

To assess the distribution of the generated images, FID has been assessed in different configurations. In the traditional UI2IT setting, one is expected to translate images from source to target domain and compare the distributions of generated images into the target domain and images that originally belong to the target domain. In this setting, with the introduction of the meta-domain corresponding to TCGA images, it is possible to observe that the distributions of images that are normalized from both source and target classification domains are closer if compared to distributions obtained by performing only the normalization of the classification source domain. These quantitative results are reported in Table 4.2 and Figure 4.8.

The PPQ metric has been used to assess the GAN-generated image quality from a pathologist's perspective. It has been calculated considering source image tiles and the corresponding generated normalized images. In this way, the pathologist can assess not only the reality of the generated image but also its consistency with respect to the original one and the lack of artifacts. Quantitative results are portrayed in Table 4.3 and Figure 4.9.

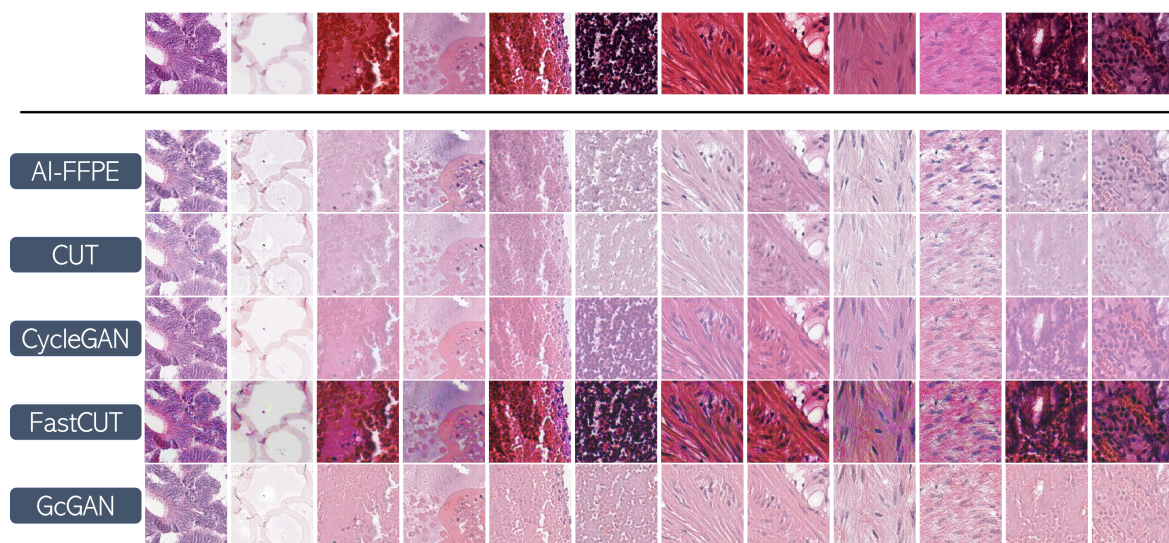


Fig. 4.4 *Normalization examples with the various UI2IT GANs considered.*

4.1.2.2 Multi-class Classification

The accuracy of the multi-class classification models has been assessed both with internal cross-validation on the CDT dataset and with external validation on the locally collected CDV

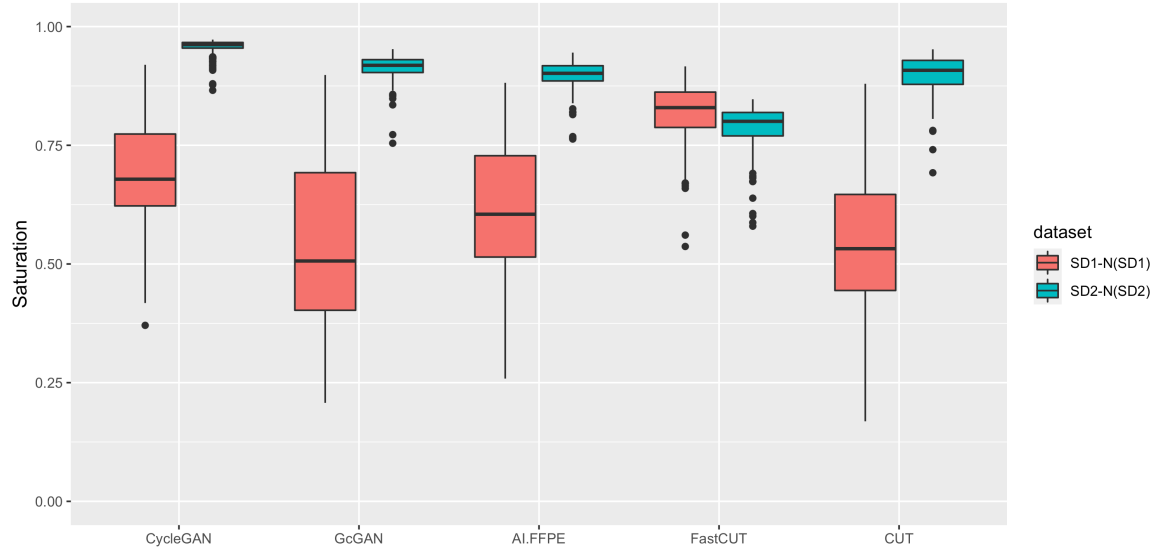


Fig. 4.5 Boxplots for SSIM between saturation channel of normalized images versus original one.

Table 4.1 Mean and standard deviation of PSNR, SSIM, LPIPS for $N(SD1)$ vs $SD1$ and $N(SD2)$ vs $SD2$.

Normalization	Data	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS VGG $\downarrow$
Reinhard	$N(SD1)$ vs $SD1$	19.8302 ± 4.2694	0.9462 ± 0.0462	0.0911 ± 0.0565
	$N(SD2)$ vs $SD2$	14.6660 ± 2.6482	0.8992 ± 0.0702	0.1333 ± 0.0400
Macenko	$N(SD1)$ vs $SD1$	22.7494 ± 5.1671	0.9542 ± 0.0516	0.0615 ± 0.0474
	$N(SD2)$ vs $SD2$	16.5445 ± 3.3880	0.8491 ± 0.1449	0.1690 ± 0.1936
Khan	$N(SD1)$ vs $SD1$	18.4453 ± 2.8344	0.6564 ± 0.1403	0.3028 ± 0.0761
	$N(SD2)$ vs $SD2$	20.3537 ± 1.0981	0.6019 ± 0.1292	0.3424 ± 0.0542
Vahadane	$N(SD1)$ vs $SD1$	21.5565 ± 4.4068	0.9397 ± 0.0637	0.0875 ± 0.0575
	$N(SD2)$ vs $SD2$	18.3096 ± 2.4472	0.8875 ± 0.0747	0.1115 ± 0.0356
CycleGAN	$N(SD1)$ vs $SD1$	16.0873 ± 3.1129	0.8449 ± 0.0508	0.1767 ± 0.0270
	$N(SD2)$ vs $SD2$	35.1468 ± 1.7279	0.9801 ± 0.0067	0.0455 ± 0.0127
GcGAN	$N(SD1)$ vs $SD1$	15.9142 ± 4.7748	0.8730 ± 0.0727	0.1731 ± 0.0487
	$N(SD2)$ vs $SD2$	32.3358 ± 1.9626	0.9783 ± 0.0097	0.0577 ± 0.0172
AI-FFPE	$N(SD1)$ vs $SD1$	16.2018 ± 3.5457	0.6255 ± 0.0672	0.3137 ± 0.0335
	$N(SD2)$ vs $SD2$	25.1388 ± 1.4350	0.8019 ± 0.0350	0.2436 ± 0.0350
FastCUT	$N(SD1)$ vs $SD1$	27.6084 ± 1.4109	0.9083 ± 0.0227	0.1874 ± 0.0336
	$N(SD2)$ vs $SD2$	20.8744 ± 1.1850	0.9130 ± 0.0163	0.1256 ± 0.0286
CUT	$N(SD1)$ vs $SD1$	14.3408 ± 3.2684	0.5837 ± 0.0730	0.3258 ± 0.0366
	$N(SD2)$ vs $SD2$	25.8854 ± 1.9922	0.8157 ± 0.0406	0.2398 ± 0.0277

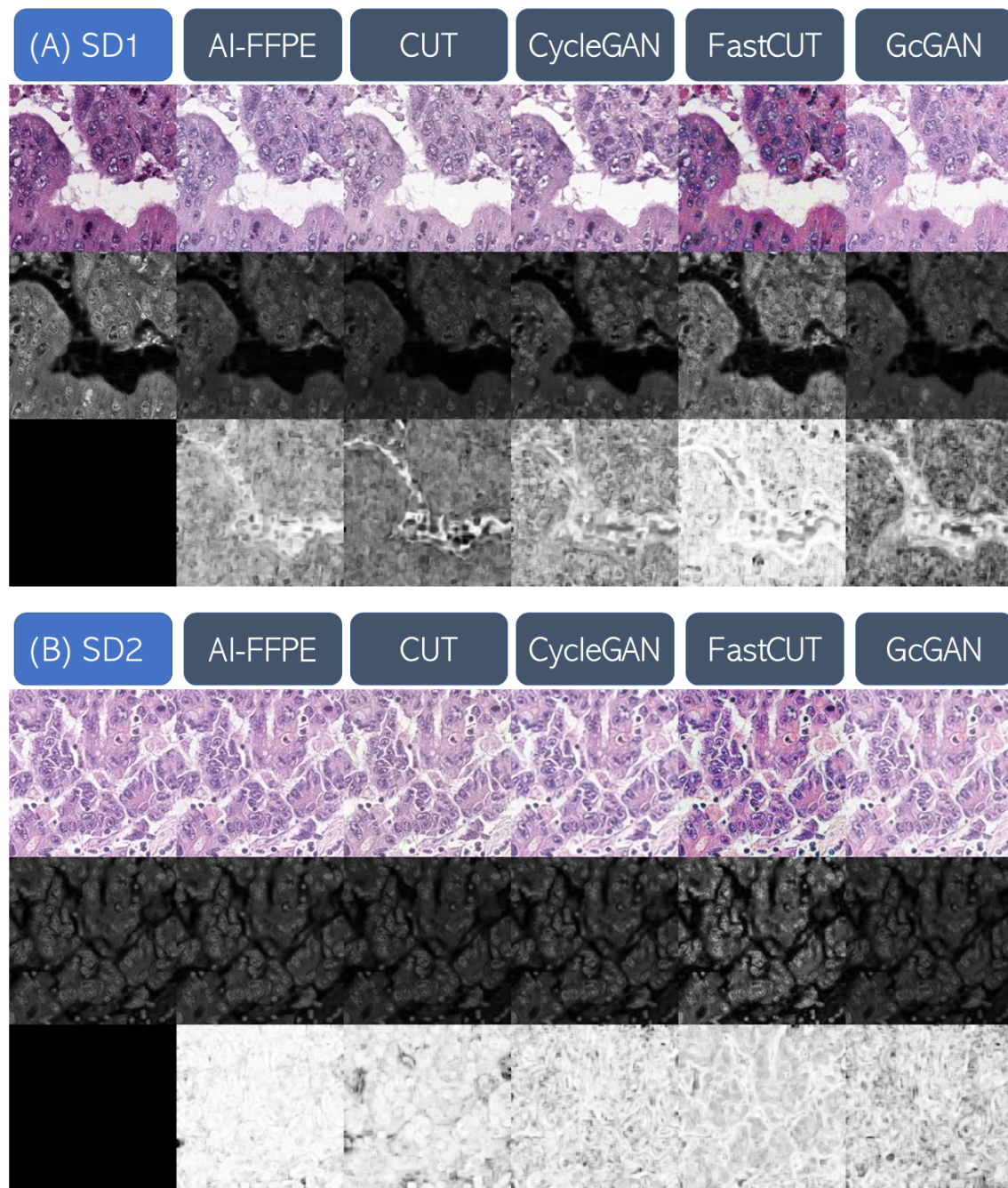


Fig. 4.6 *Saturation differences between the considered UI2IT methods.* (A) Tile from SD1 dataset. (B) Tile from SD2 dataset. (First Row) Original tile in the first column, then different GAN-based normalization methods. (Second Row) Saturation channel of the images in the first row. (Third Row) Difference between the saturation of normalized image compared to the original one.

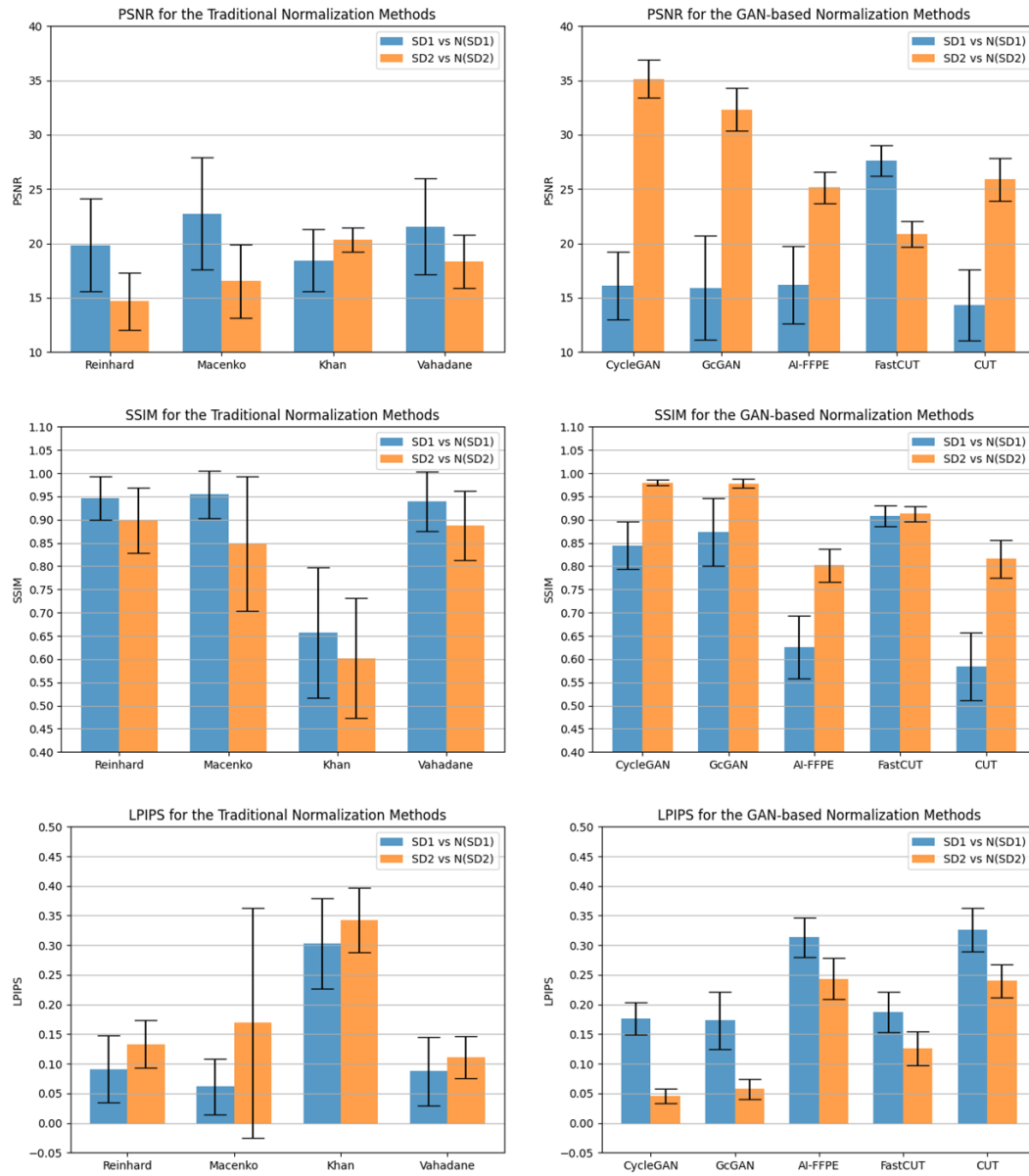


Fig. 4.7 Bar Plots with error bars of PSNR, SSIM, LPIPS for $N(SD1)$ vs $SD1$ and $N(SD2)$ vs $SD2$.

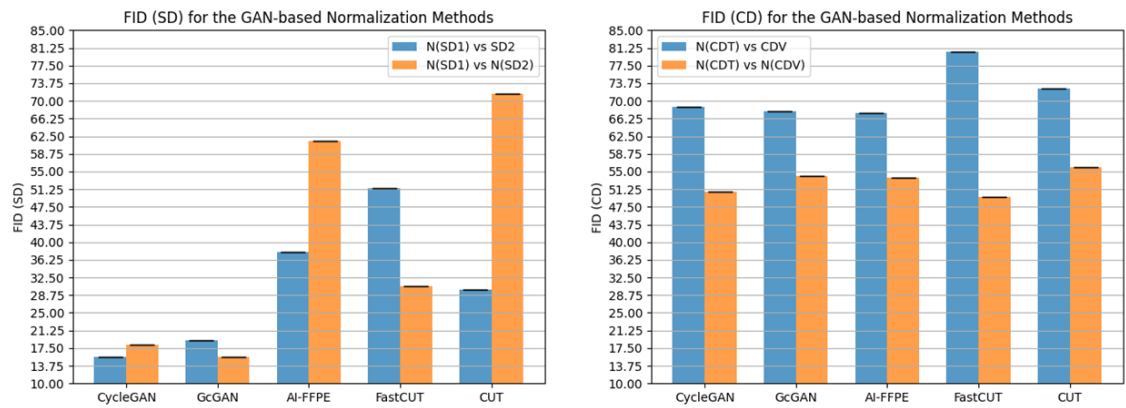


Fig. 4.8 Bar Plots with error bars of FID for $N(SD1)$ vs $SD2$, $N(SD1)$ vs $N(SD2)$, $N(CDT)$ vs CDV , and $N(CDT)$ vs $N(CDV)$.

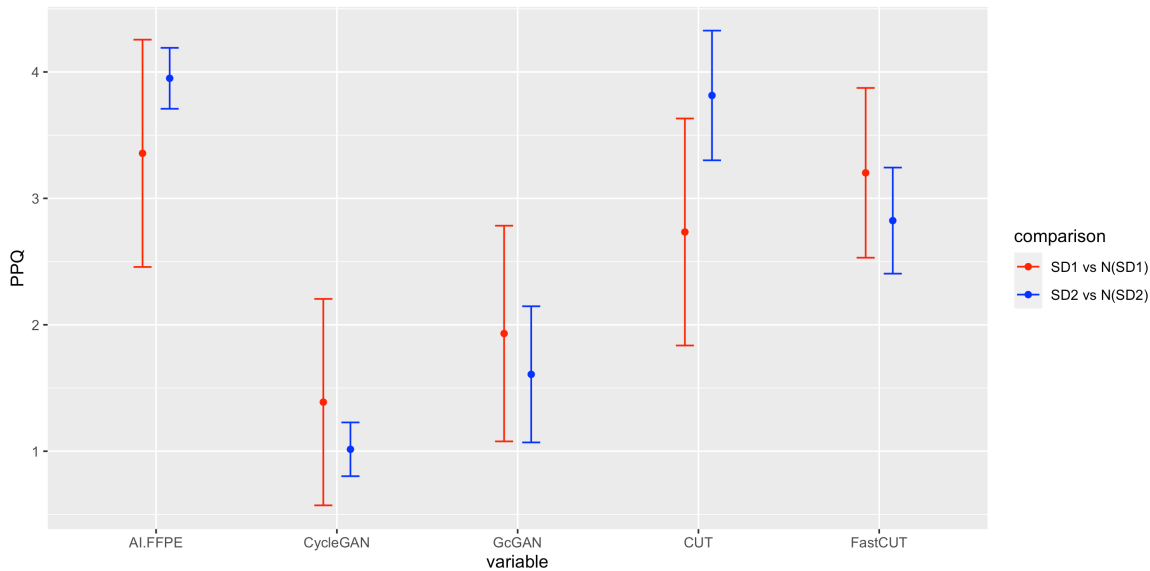


Fig. 4.9 Line Plots of PPQ for $SD1$ vs $N(SD1)$ and $SD2$ vs $N(SD2)$.

Table 4.2 *FID comparison for the various normalization methods.*

Normalization	Data	FID ↓
None	SD1 vs SD2	35.9791
	CD1 vs CD2	73.8733
Reinhard	N(SD1) vs N(SD2)	26.6528
	N(CD1) vs N(CD2)	58.1405
Macenko	N(SD1) vs N(SD2)	26.6980
	N(CD1) vs N(CD2)	53.7302
Khan	N(SD1) vs N(SD2)	22.9181
	N(CD1) vs N(CD2)	44.1937
Vahadane	N(SD1) vs N(SD2)	26.6518
	N(CD1) vs N(CD2)	56.6086
CycleGAN	N(SD1) vs SD2	15.5333
	N(SD1) vs N(SD2)	18.2566
	N(CD1) vs CD2	68.6784
	N(CD1) vs N(CD2)	50.7333
GcGAN	N(SD1) vs SD2	19.2417
	N(SD1) vs N(SD2)	15.6546
	N(CD1) vs CD2	67.8117
	N(CD1) vs N(CD2)	54.1416
AI-FFPE	N(SD1) vs SD2	37.8711
	N(SD1) vs N(SD2)	61.5486
	N(CD1) vs CD2	67.3626
	N(CD1) vs N(CD2)	53.6819
FastCUT	N(SD1) vs SD2	51.5162
	N(SD1) vs N(SD2)	30.5891
	N(CD1) vs CD2	80.4216
	N(CD1) vs N(CD2)	49.5950
CUT	N(SD1) vs SD2	29.9900
	N(SD1) vs N(SD2)	71.4490
	N(CD1) vs CD2	72.5915
	N(CD1) vs N(CD2)	55.9412

Table 4.3 *Mean and standard deviation of PPQ for N(SD1) vs SD1 and N(SD2) vs SD2.*

Normalization Method	Data	PPQ ↓
CycleGAN	N(SD1) vs SD1	1.3883 ± 0.8165
	N(SD2) vs SD2	1.0151 ± 0.2127
GcGAN	N(SD1) vs SD1	1.9309 ± 0.8531
	N(SD2) vs SD2	1.6080 ± 0.5385
AI-FFPE	N(SD1) vs SD1	3.3564 ± 0.8990
	N(SD2) vs SD2	3.9497 ± 0.2410
FastCUT	N(SD1) vs SD1	3.2021 ± 0.6715
	N(SD2) vs SD2	2.8241 ± 0.4195
CUT	N(SD1) vs SD1	2.7340 ± 0.8977
	N(SD2) vs SD2	3.8141 ± 0.5131

dataset. These quantitative results are shown in Figure 4.10 and Table 4.4. Other classification metrics are reported in Table 4.5. The DenseNet201 model consistently outperforms the other deep feature extractors in all the scenarios. Among the traditional normalization methods, Reinhard is the one that produces the highest validation accuracy. Among the GAN-based normalization methods, FastCUT achieves slightly better performances than the other methods. FastCUT also presents better performance than the Reinhard method. The quality of the features has been assessed by looking at embedding plots for the normalization methods. The embedding plot associated to the features extracted by DenseNet201, after FastCUT normalization, can be seen from Figure 4.11.

4.1.2.3 Relationships between Normalization and Classification

Several interesting relationships have been observed from the analysis presented in this section. Indeed, as it can be observed from Figure 4.12, there is a trend between the performance of the classification methods and the quality of the distribution of the generated images. This trend is visible for all the considered metrics, but it is statistically significant only for Top3-Accuracy (VGG16 achieves $r=0.9446$, $p=0.0155$), mDice (VGG16 achieves $r=0.9075$, $p=0.0333$), and mPrecision (DenseNet201 achieves $r=0.9950$, $p=0.0004$ and InceptionV3 achieves $r=0.9495$, $p=0.0135$). FID has been normalized by taking its reciprocal, so to make it a measure in which higher values correspond to better quality distributions.

The PPQ has also been found correlated to FID and to LPIPS, as can be seen from Figure 4.13. These results are also statistically significant (PPQ SD1 vs FID: $r=0.8968$, $p=0.0392$; PPQ SD2 vs LPIPS SD2: $r=0.9775$, $p=0.0040$).

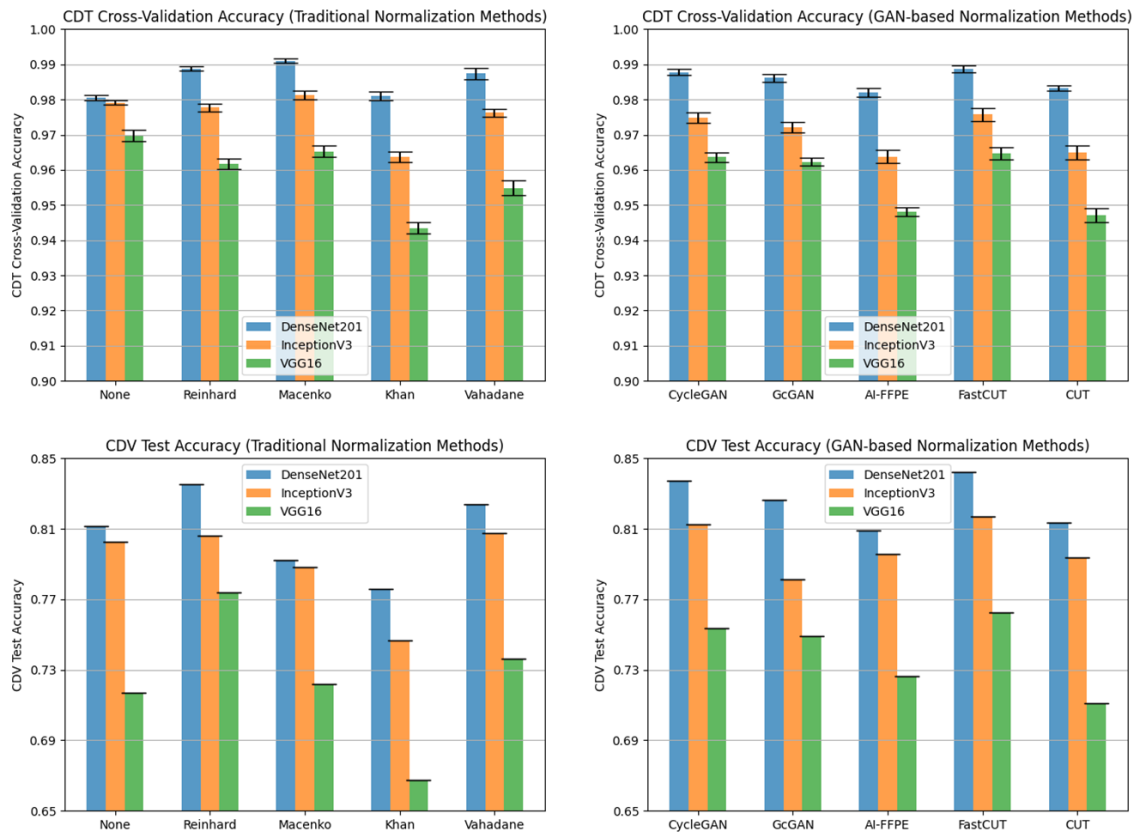


Fig. 4.10 Accuracy for the various normalization methods on the CDT and CDV datasets.

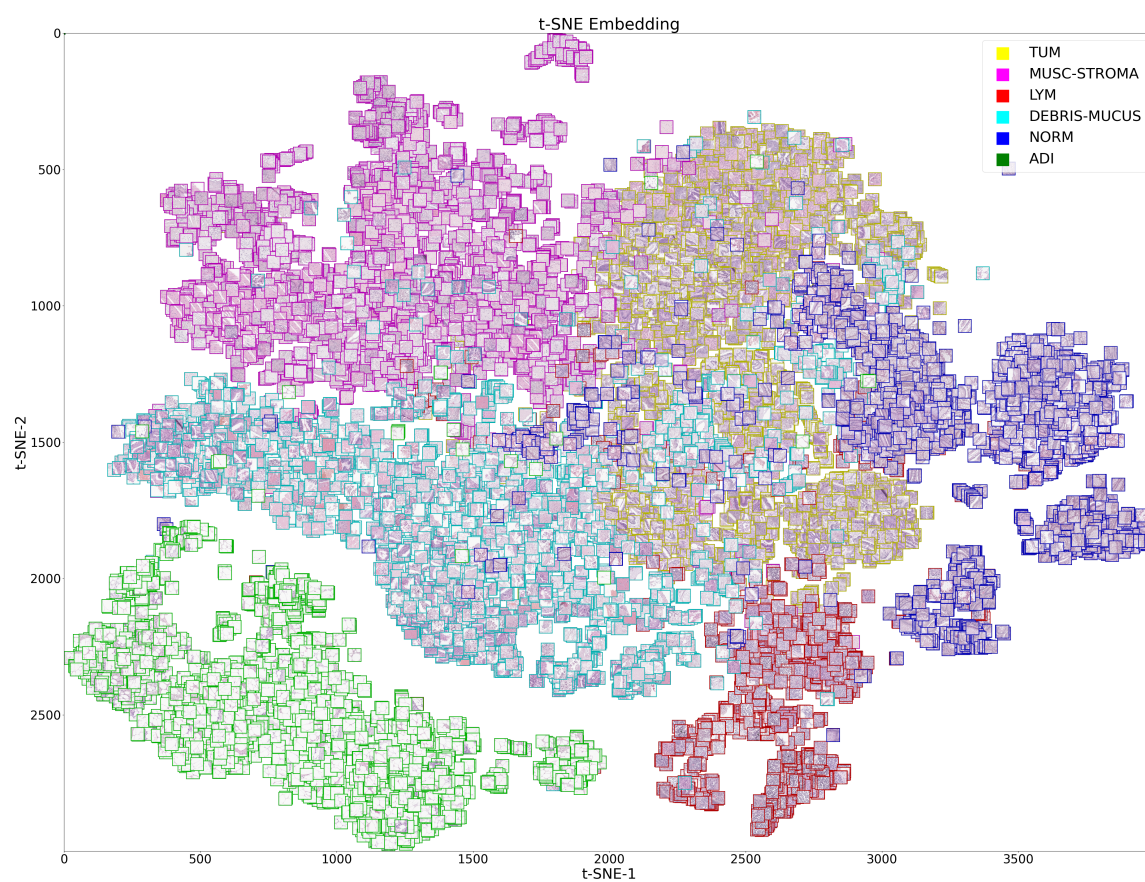


Fig. 4.11 Embedding plot obtained applying *t*-SNE to features extracted from DenseNet201 on the CDV dataset, with the FastCUT normalization methodology.

Table 4.4 Accuracy on the CDT and CDV classification datasets.

Normalization	Architecture	CDT	CDV		
		Accuracy	Top 1-Accuracy	Top 2-Accuracy	Top 3-Accuracy
None	InceptionV3	0.9791 (0.0006)	0.8025	0.9209	0.9608
	DenseNet201	0.9805 (0.0008)	0.8115	0.9253	0.9721
	VGG16	0.9697 (0.0016)	0.7168	0.8688	0.9376
Reinhard	InceptionV3	0.9777 (0.0011)	0.8063	0.9133	0.9586
	DenseNet201	0.9888 (0.0007)	0.8352	0.9261	0.9652
	VGG16	0.9617 (0.0014)	0.7737	0.8964	0.9473
Macenko	InceptionV3	0.9813 (0.0012)	0.7884	0.8978	0.9486
	DenseNet201	0.9910 (0.0006)	0.7924	0.8969	0.9557
	VGG16	0.9653 (0.0015)	0.7219	0.8701	0.9244
Khan	InceptionV3	0.9637 (0.0016)	0.7465	0.8871	0.9449
	DenseNet201	0.9811 (0.0012)	0.7757	0.9070	0.9560
	VGG16	0.9435 (0.0016)	0.6673	0.8319	0.9168
Vahadane	InceptionV3	0.9762 (0.0011)	0.8075	0.9179	0.9603
	DenseNet201	0.9874 (0.0016)	0.8240	0.9265	0.9646
	VGG16	0.9548 (0.0021)	0.7362	0.8787	0.9390
CycleGAN	InceptionV3	0.9749 (0.0015)	0.8126	0.9256	0.9678
	DenseNet201	0.9878 (0.0009)	0.8375	0.9413	0.9793
	VGG16	0.9636 (0.0013)	0.7538	0.8923	0.9461
GcGAN	InceptionV3	0.9720 (0.0015)	0.7811	0.9161	0.9657
	DenseNet201	0.9861 (0.0011)	0.8266	0.9363	0.9747
	VGG16	0.9623 (0.0011)	0.7494	0.8841	0.9410
AI-FFPE	InceptionV3	0.9638 (0.0018)	0.7957	0.9062	0.9536
	DenseNet201	0.9820 (0.0013)	0.8090	0.9217	0.9693
	VGG16	0.9481 (0.0012)	0.7263	0.8662	0.9355
FastCUT	InceptionV3	0.9757 (0.0019)	0.8170	0.9305	0.9702
	DenseNet201	0.9886 (0.0010)	0.8420	0.9410	0.9774
	VGG16	0.9646 (0.0017)	0.7626	0.8956	0.9516
CUT	InceptionV3	0.9649 (0.0020)	0.7935	0.9127	0.9615
	DenseNet201	0.9833 (0.0008)	0.8134	0.9262	0.9728
	VGG16	0.9471 (0.0020)	0.7111	0.8517	0.9332

Table 4.5 *Classification Mean Precision, Recall, and Dice (mPrecision, mRecall, and mDice, respectively).*

Normalization	Architecture	CDV		
		mPrecision	mRecall	mDice
None	InceptionV3	0.8361	0.7668	0.7835
	DenseNet201	0.8479	0.7920	0.8039
	VGG16	0.8007	0.6743	0.6967
Reinhard	InceptionV3	0.8354	0.7780	0.7902
	DenseNet201	0.8643	0.8038	0.8172
	VGG16	0.8268	0.7278	0.7492
Macenko	InceptionV3	0.8171	0.7594	0.7722
	DenseNet201	0.8291	0.7740	0.7874
	VGG16	0.8291	0.6778	0.7203
Khan	InceptionV3	0.7893	0.7092	0.7263
	DenseNet201	0.8249	0.7397	0.7623
	VGG16	0.7289	0.6070	0.6218
Vahadane	InceptionV3	0.8369	0.7801	0.7948
	DenseNet201	0.8605	0.7955	0.8124
	VGG16	0.8219	0.6890	0.7226
CycleGAN	InceptionV3	0.8543	0.7754	0.7952
	DenseNet201	0.8761	0.8151	0.8349
	VGG16	0.8263	0.7153	0.7384
GcGAN	InceptionV3	0.8399	0.7439	0.7684
	DenseNet201	0.8645	0.8030	0.8229
	VGG16	0.8187	0.7130	0.7324
AI-FFPE	InceptionV3	0.8470	0.7542	0.7754
	DenseNet201	0.8664	0.7788	0.8003
	VGG16	0.8105	0.6927	0.7158
FastCUT	InceptionV3	0.8667	0.7784	0.8017
	DenseNet201	0.8800	0.8170	0.8372
	VGG16	0.8069	0.7315	0.7522
CUT	InceptionV3	0.8397	0.7614	0.7823
	DenseNet201	0.8616	0.7858	0.8067
	VGG16	0.7829	0.6865	0.7085

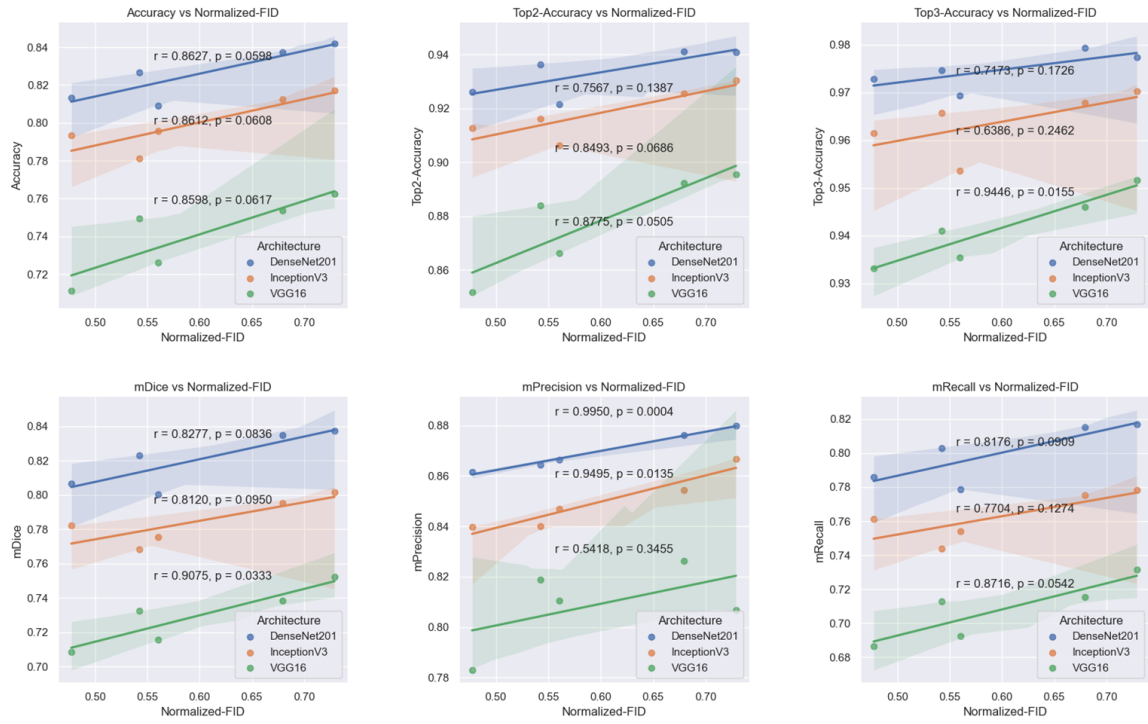


Fig. 4.12 Correlations between classification metrics and FID.

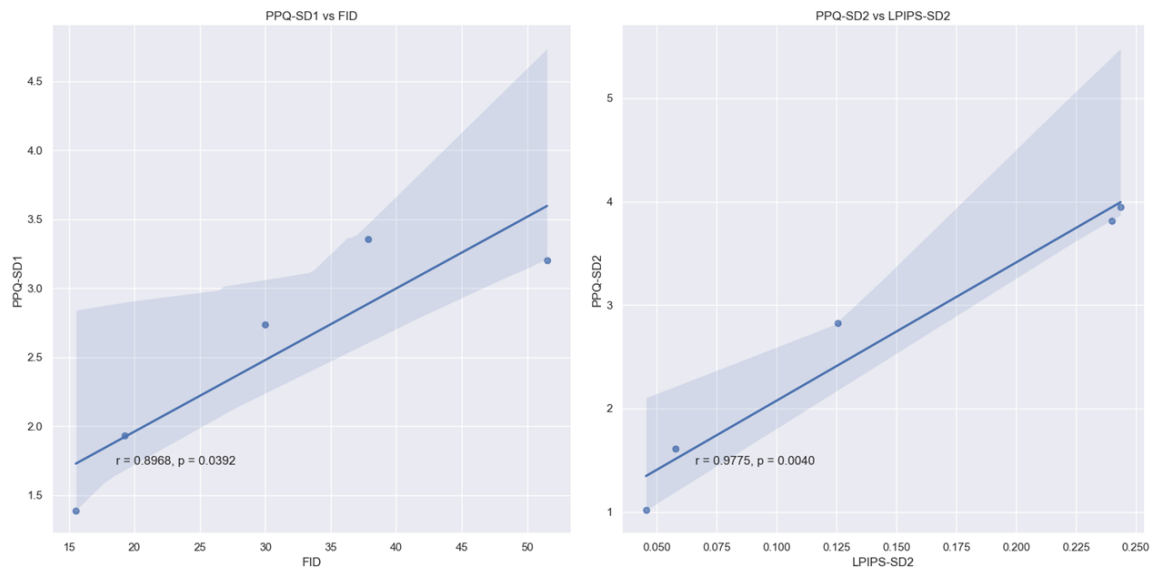


Fig. 4.13 Correlation between FID and PPQ, and between LPIPS and PPQ.

4.1.3 Discussion

In this study, several GAN-based UI2IT methods have been characterized with the aim to perform stain color normalization for histopathological tissue classification. Contrary to what is usually done for GAN-based normalization, that is, considering only source and target domains, a meta-domain has been considered, which in this case consists of WSIs belonging to the TCGA. This meta-domain, containing tissues from a wide variety of laboratories, can allow learning a mapping to the considered target domain that is more general than previous research works, thus avoiding the need to train a normalization model for every pair of possible data domains for histopathological data. Due to this difference, the proposed methodology requires performing double normalization at inference time (that is, performing both $G_{MT}(S)$ and $G_{MT}(T)$, instead of only $G_{MT}(S)$), which leads to an improvement in the FID among the data distributions of the classification dataset. Indeed, from Figure 4.8 it is possible to observe that FID is consistently lower for all the considered GAN models when normalization is performed to both CDT and CDV, and not only CDT. This may look counterintuitive since the learned mapping has already WSIs of the target domain coming from the same distribution as CDV. Instead, when considering the stain color normalization datasets, it is worth noting that performing double normalization is not always convenient. Indeed, SD1 and SD2 are the source domain and target domain, respectively, selected for learning the GAN mapping. In this case, CycleGAN, AI-FFPE, and CUT show worse FID results when double normalization is applied instead of single normalization.

In this study, PPQ, a perceptive quality metric based on the assessment performed by the pathologist on the GAN-generated image tiles, has been introduced. As can be seen from Figure 4.9, CycleGAN and GcGAN were the two best-performing methods according to PPQ. This result is perfectly consistent with: (i) FID values observed on N(SD1) vs SD2 and N(SD1) vs N(SD2); (ii) LPIPS values observed on SD1 vs N(SD1) and SD2 vs N(SD2). Therefore, this study can confirm that CNN-features-based metrics to assess the perceptive quality of generated images can be exploited also in histopathological contexts. As can be seen from Figure 4.13, FID and PPQ, and LPIPS and PPQ, are also correlated when considering the GAN-based normalization results.

The effectiveness of CycleGAN and GcGAN images is also observed by the fact that they possess the highest values for PSNR and SSIM for what concerns SD2 vs N(SD2), as observed in Figure 4.7. This means that those methods are the most suitable for performing transformation of the target domain images with $G_{MT}(T)$, which is useful as previously mentioned (considering the increase in FID for double normalization).

With respect to the multi-class classification task, the DenseNet201 model consistently outperforms InceptionV3 and VGG16, in all considered normalization paradigms, as portrayed in Figure 4.10. The normalization methodology which permits achieving the best performance on the locally collected external validation set is FastCUT. Though FastCUT does not achieve satisfactory results in reality of generated image tiles, since it tends to over-saturate the input images (quantitative assessment, see Figure 4.5; visual examples, see Figure 4.6), when applied in double normalization configuration, it can still offer a useful normalization for the classification task. This can be specially noted by comparing the FID of N(SD1) vs SD2 with FID of N(SD1) vs N(SD2), and FID of N(CDT) vs CDV with FID of N(CDT) vs N(CDV). Indeed, after double normalization is performed, FastCUT achieves the best FID on the classification datasets. This result means that the perceptive quality of transformed images is not necessarily correlated to better classification accuracies of downstream classifiers. Though, the perceptive quality of generated images has been considered a useful asset for the introduction of GAN-based normalization models in the clinical routine.

The quality of the features extracted by DenseNet201 after having performed normalization with FastCUT can be seen in Figure 4.11. The different tissue classes appear clearly clustered in the 2D embedding plot obtained with t-SNE.

Classification results are correlated to the quality of image distributions, as can be seen from correlations between normalized FID and classification performance reported in Figure 4.12. This confirms that FID or similar quality metrics should be checked when performing GAN-based normalization (i.e., conditional generation) since they can already offer important insight into how the models adopted for the downstream tasks will perform.

4.1.4 Summary

In this work, three CNN architectures and nine normalization techniques have been considered for the sake of realizing a pipeline that is robust to stain color variation in a CRC histological classification setup.

CRC multi-class tissue classification is an important task in digital pathology. In particular, to date, the study of tumors is moving toward the integration of genomic data, such as transcriptomics and its spatial localization. In the work presented in this section, the focus is on CRC, but the described approach could be considered as a proof of concept for other malignancies. Indeed, the segmentation task could be a preliminary step in digital pathology

studies, for instance, to study the relationship between the tumor, its microenvironment, and genomic features.

Since the preparation of histological slides is a complex process, composed of various stages, and differences can be introduced in any of them, color normalization is a fundamental step to effectively perform quantitative tasks.

Traditional normalization methods are easy-to-use, since they only need a template tile, but generate images that are subject to color artifact, making them unsuitable for subsequent pathologists' analyses. On the other side, the realm of GAN architectures to implement UI2IT can offer a powerful framework to carry out stain color normalization, allowing the achievement of impressive results both from the perspective of quality and reality of generated tiles and for the performance of the models involved in the subsequent tasks.

Nonetheless, care has to be reserved for assessing GAN-generated image tiles, and several quality measures should be included to have a complete overview of which model may work better for the task under consideration.

Observations of pathologists may help in assessing the quality of generated images. Indeed, they have been included in this study with the PPQ metric, which can be used to have a medical expert view that can eventually confirm the quality of the distributions measured with quantitative measures such as FID.

4.2 Nuclei Detection from Histopathological Images

The literature review (Section 2.3.2) for nuclei detection showed that methodologies prior to the advent of deep learning demonstrate worse performance on this task. Moreover, handcrafted feature extraction is a tedious and complex process, which can lead to different results depending on the experience of the feature engineers and domain experts. It is worth noting that CNN-based approaches require datasets with a distinct label for every nucleus, based on observations made in the last several years. Simple existing semantic segmentation methods, trained without the knowledge of different instances, cannot be reliably adopted for nuclei detection.

Many cell nuclei detection methods share a basic approach that includes generating an intermediate map through a CNN that indicates the presence of a nucleus, called the probability or proximity map (P-Map) [23, 434], or have specialized architectures that are trained to individuate the centers of the nuclei, such as SC-CNN [347]. Indeed, the P-Map represents proximities as a monochromatic image: the intensities have high values near the centroid of the nucleus, and gradually lower going toward the boundaries.

By following the idea of determining a structure similar to a P-Map, a novel method for nuclei detection has been proposed, without the need for specialized architectures or hand-crafted feature extraction; rather, only semantic segmentation networks and XAI techniques have been used. The proposed method is quick to train and is extensible, considering that it can be plugged on top of existing semantic segmentation networks.

The presence of clustered or overlapped nuclei with semantic segmentation models can be spotted on visual inspection of the images. In order to overcome this issue, the potentialities of the Grad-CAM method for segmentation have been exploited, which made it possible to highlight the activation of the nucleus class (compared to the background class), thus obtaining a saliency map with properties similar to the classic P-Map. The locations of the nuclei are subsequently determined by looking for local maxima in the activation map. Starting from the identified centroids, it is possible to associate all the pixels belonging to the considered nucleus, with a proximity criterion. This model alone, which will be denoted as Nuclei Detection with Grad-CAM (NDG-CAM), was capable of achieving performance in line with state-of-the-art methods. Because the Mask R-CNN [93] instance segmentation architecture is widely employed and constitutes a standard baseline for these tasks, a combined model for further enhancing the results has also been realized, surpassing the state-of-the-art.

To summarize, the contributions of this section can be considered as follows: (i) the introduction of a novel detection method for nuclei—NDG-CAM—which exploits Grad-CAM for semantic segmentation; (ii) the collection of a local annotated dataset of patients diagnosed with colorectal cancer to show the applicability of the proposed method in a local hospital; (iii) the comparison of the proposed method with different state-of-the-art techniques to show its effectiveness; (iv) the training and evaluation of an instance segmentation architecture as the baseline; and (v) the design of a combined model which, exploiting both NDG-CAM and Mask R-CNN, can surpass the current literature performance concerning nuclei detection.

The remainder of the section is organized as follows. Section 4.2.1 first describes the datasets adopted for the analysis. Then, semantic segmentation configurations and architectures are presented. The NDG-CAM is proposed, and its workflow is delineated. An instance segmentation architecture is also considered as the baseline. Lastly, implementation details, the combined model, and the evaluation metrics employed for the analysis are presented. Results are portrayed in Section 4.2.2 and discussed in Section 4.2.2.1. A comparison with other state-of-the-art approaches is considered here. Lastly, final remarks, conclusions, and ideas for future works are drawn in Section 4.2.3.

4.2.1 Materials and Methods

4.2.1.1 Datasets

For the tasks of nuclei segmentation and detection, different datasets were considered in order to find the best-performing model. In particular, the latest and largest publicly available datasets for nuclei detection and segmentation were considered. Moreover, a local dataset has been collected, to prove the feasibility of the proposed system on new data from a local hospital.

- **MoNuSeg** [338, 435, 436]. The cell nucleus segmentation dataset used in this work is publicly accessible from the 2018 Data Science Bowl challenge [437]. The dataset contains a large number of segmented nuclei images and includes different cell types; there are 30 training H&E images containing 21,623 hand-annotated nuclear boundaries from the breast, kidney, prostate, liver, colon, bladder, and stomach. Moreover, there are also 14 H&E test images containing 7000 nuclear boundary annotations from the breast, kidney, prostate, colon, bladder, lung, and brain. All images, each of size 1000×1000 , were captured at $40\times$ magnification. The nuclear contour annotations are provided through XML files.
- **CRCHistoPhenotypes**: Labeled Cell Nuclei Data [347, 438]. This publicly available dataset contains 100 H&E-stained histology images of colon cell nuclei obtained from WSI of 10 patients with a magnification factor of $20\times$. Tiles have a size of 500×500 . Nuclear annotations are provided through the coordinates of the centroids in *.mat* format, resulting in a total of 29,756 annotated nuclei for detection purposes.
- **NuCLS** [439]. The dataset contains over 220,000 labeled nuclei from breast cancer images from TCGA, obtained from 125 patients with breast cancer (1 slide per patient) and captured with a magnification factor of $40\times$. These nuclei were annotated through the collaborative effort of pathologists, pathology residents, and medical students. Data from both single-rater and multi-rater studies are provided. For single-rater data, there are both pathologist-reviewed and uncorrected annotations. For multi-rater datasets, there are annotations generated with and without suggestions from weak segmentation and classification algorithms. Only the single-rater dataset was used for this study, which is already split into train and test sets. The annotations for the single-rater dataset include 59,485 nuclei and 19,680 boundaries, extracted from 1744 H&E image tiles of variable dimensions between 200 and 400 pixels.

- **Local dataset** from Pathology Department of IRCCS Istituto Tumori "Giovanni Paolo II" [20]. This cohort consists of 19 H&E image tiles which overall contain more than 6378 nuclei from patients with colorectal cancer. Images have a size of 512×512 and were captured at 40 $\times$ magnification. Annotations have been provided by a biologist with experience in analyzing histopathological data.

Hereafter, the training and test sets of MoNuSeg (D1) will be denoted with T1 and V1. The overall dataset of CRCHistoPhenotypes, instead, will be referred to as D2. The Mask R-CNN model has been trained on the NuCLS (D3) dataset, being the largest publicly available dataset with annotations formatted for instance segmentation. Because D1 already includes a validation set, it has been exploited for the first validation stage. As an independent external validation set, other image tiles from the Pathology Department of IRCCS Istituto Tumori "Giovanni Paolo II" [20] have been collected. This dataset will be denoted as V4 and has the purpose to evaluate the generalization capability of the best semantic segmentation network configuration individuated with the D1 and D2 datasets, and the Mask R-CNN model trained on the D3 dataset. Figure 4.14 summarizes the pipeline implemented for training and validating the models.

A summary of the details for the employed datasets is reported in Table 4.6, whereas sample images are depicted in Figure 4.15.

Table 4.6 *Summary of datasets for nuclei.*

Dataset	Publication Year	Organs	Resolution	Number of H&E images	Number of Nuclei	Size (pixels)	Annotations Format
MoNuSeg—Train (T1) [338]	2017	breast, kidney, prostate, liver, colon, bladder, stomach	40 $\times$	30	21,623	1000 $\times$ 1000	Nuclei Contours
MoNuSeg—Test (V1) [338]		breast, kidney, prostate, colon, bladder, lung, brain		14	7000		
CRCHistoPhenotypes (D2) [347]	2016	colon	20 $\times$	100	29,756	500 $\times$ 500	Nuclei Centroids
NuCLS (D3) [439]	2019	breast	40 $\times$	1744	59,485	200–400 per side	Nuclei Contours or Bounding Boxes
Local (V4)	2022	colon	40 $\times$	19	6378	512 $\times$ 512	Nuclei Centroids

4.2.1.2 NDG-CAM

In this section, the methodology adopted for NDG-CAM is introduced. Several steps have been carried out. As the first step, a semantic segmentation architecture trained for nuclear segmentation is required. Different experimental configurations of the datasets and network architectures have been compared in order to find the most suitable model, with

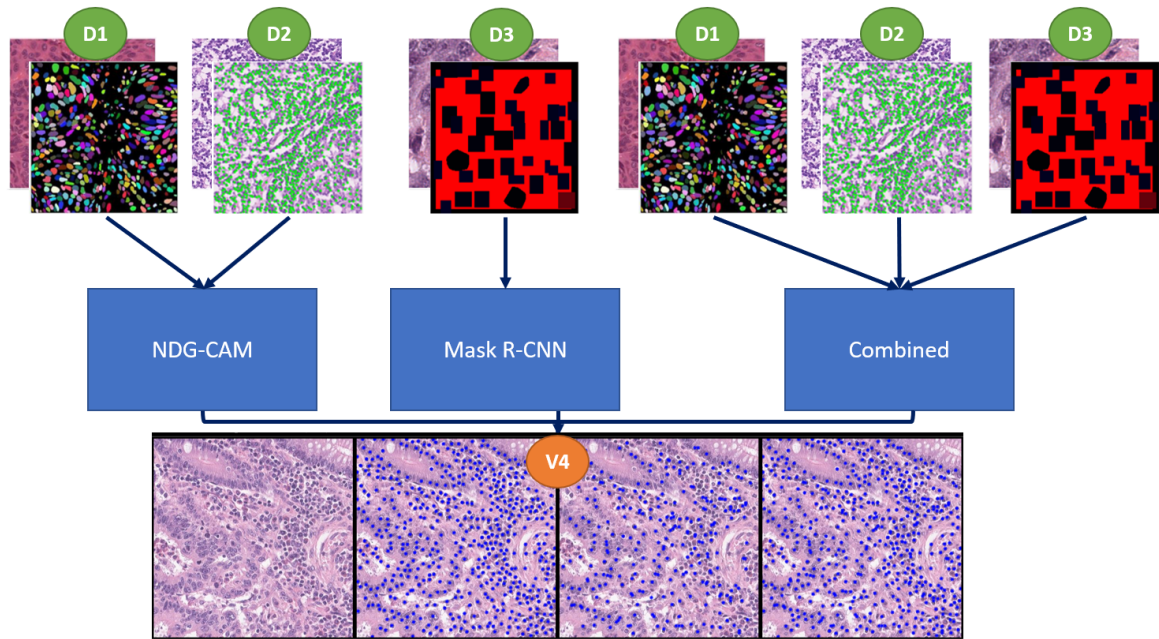


Fig. 4.14 *Pipeline adopted for training and validation.* D1 and D2 datasets have been used to train and select the best semantic segmentation network. The D3 dataset has been exploited to train the Mask R-CNN instance segmentation architecture. Finally, external validation has been conducted on the local validation dataset V4.

details reported in the following paragraphs. Then, the Grad-CAM technique for semantic segmentation, which is still underexplored if compared to Grad-CAM for classification, has been employed to obtain saliency maps of the nuclei, with higher values of intensity corresponding to positions nearest to the centroids. Subsequently, a search for local maxima, combined with post-processing and clustering, allowed for the detection and eventually instance segmentation of the nuclei. Compared to specialized architectures, such as those used for instance segmentation, semantic segmentation networks are simpler and faster to train. In addition, the proposed system can be trained if labels do not distinguish between different nuclear instances, which would not be possible for instance segmentation models.

Semantic Segmentation Workflow Starting from the datasets described in the previous sections, the following experiments were carried out, all with images at a size of 512×512 :

- a Train on D2 and validation on V1 at 20× resolution.
- b Train on T1 and validation on V1 at 20× resolution.
- c Train on T1 and validation on V1 at 40× resolution.

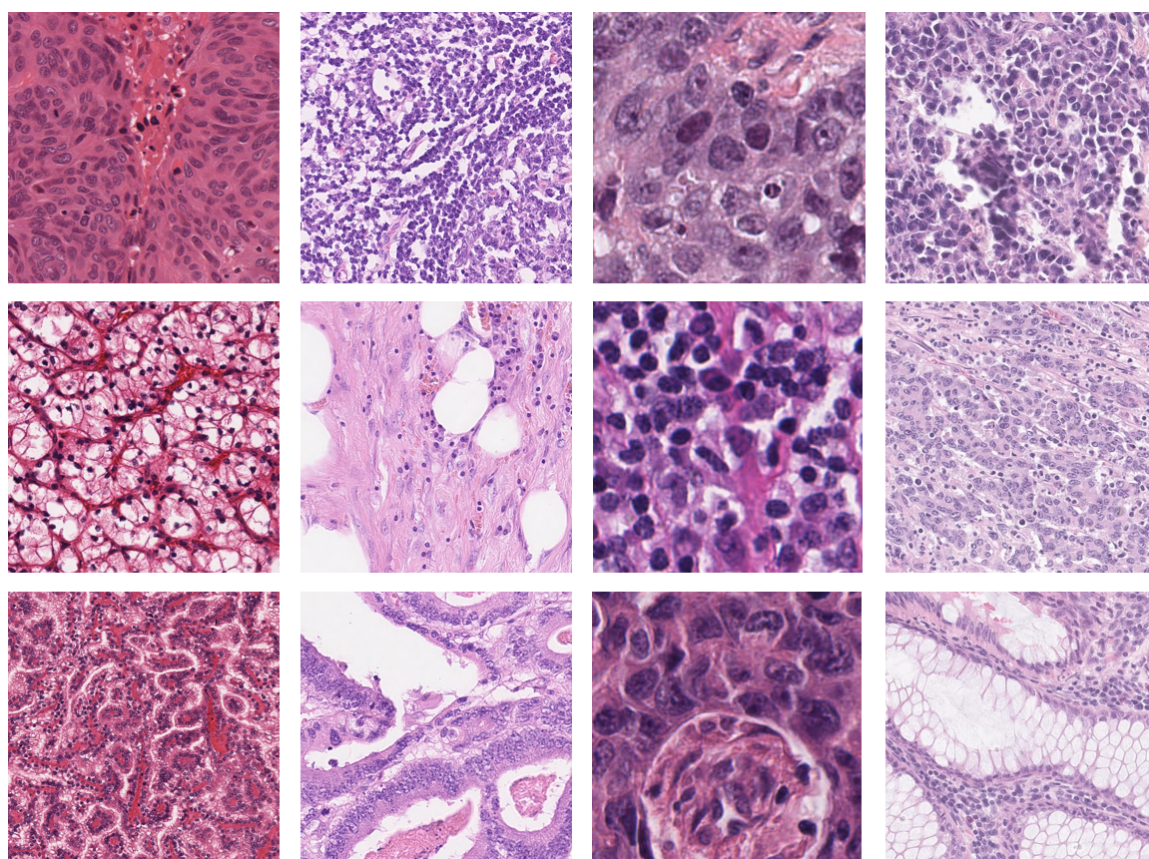


Fig. 4.15 *Sample images of datasets for nuclei detection.* (First column) D1—MoNuSeg [435]; (second column) D2—CRCHistoPhenotypes [438]; (third column) D3—NuCLS [439]; (fourth column) V4—local dataset [20].

In the first two experiments, images were padded from 500×500 to 512×512 exploiting the mirror padding. Instead, in the last experiment, the images were padded from 1000×1000 to 1024×1024 with mirror padding and subsequently divided into 4 tiles of 512×512 . For each experiment, different deep network architectures were trained and compared: U-Net [72], SegNet [70], and DeepLab v3+ [80] in three different backbone configurations, namely ResNet18, ResNet50 [48], and MobileNet-v2 [440]. The aforementioned experiments were carried out in MATLAB[®] R2021a.

Network Architectures The segmentation phase is a milestone for the detection phase; this step aims to discriminate between cell nuclei and the background. Semantic segmentation architectures play a role of pivotal importance in deep learning-based medical image analysis [16, 35, 37, 56]. It is a process that associates a label or a category to each pixel of an input image, thus allowing the pixelwise spatial localization of each object category appearing in the scene.

In the specific case under analysis, the goal was to segment the cell nuclei in a robust way, so as to provide satisfactory results even when the algorithm would have been applied to different images of the same type. For this reason, it was decided to carry out the same experiments with several convolutional architectures.

The considered architectures include:

- U-Net [72]. It is a fully convolutional network to perform the semantic segmentation task. The U-Net architecture consists of a series of encoding layers and contractions that are used to extract the context of the image, followed by a sequence of symmetrical decoding layers and expansions to recover the spatial information. In the exploited MATLAB[®] setting, the network is characterized by 58 convolutional layers; the first layer deals with a z-score normalization of the inputs, whereas the last one presents the Dice function as a loss function.
- SegNet [70]. This is another encoder–decoder architecture. In this case, the decoding blocks exploit max pooling indices received from the corresponding contraction block to perform the oversampling, instead of using trainable upsampling layers as transposed convolutions. In the considered MATLAB[®] setting, this CNN consists of 31 layers with a cross-entropy loss function.
- DeepLab v3+ [80]. This architecture features atrous spatial pyramid pooling (ASPP) and the encoder–decoder paradigm. The first aspect concerns a particular way of

combining layers of atrous and depthwise convolution, with which the model captures and concatenates features at different scales. For this network, the backbone is customizable. Three different basic CNN encoders were used: ResNet18, ResNet50, and MobileNet-v2. The DeepLab v3+ has 100 layers, of which the last is a softmax layer that is used to obtain the probabilities that each pixel belongs to the nucleus or background class; in this case, the chosen loss function is the Dice loss.

An example of semantic segmentation prediction from DeepLab v3+ with backbone ResNet18 is shown in Figure 4.16.

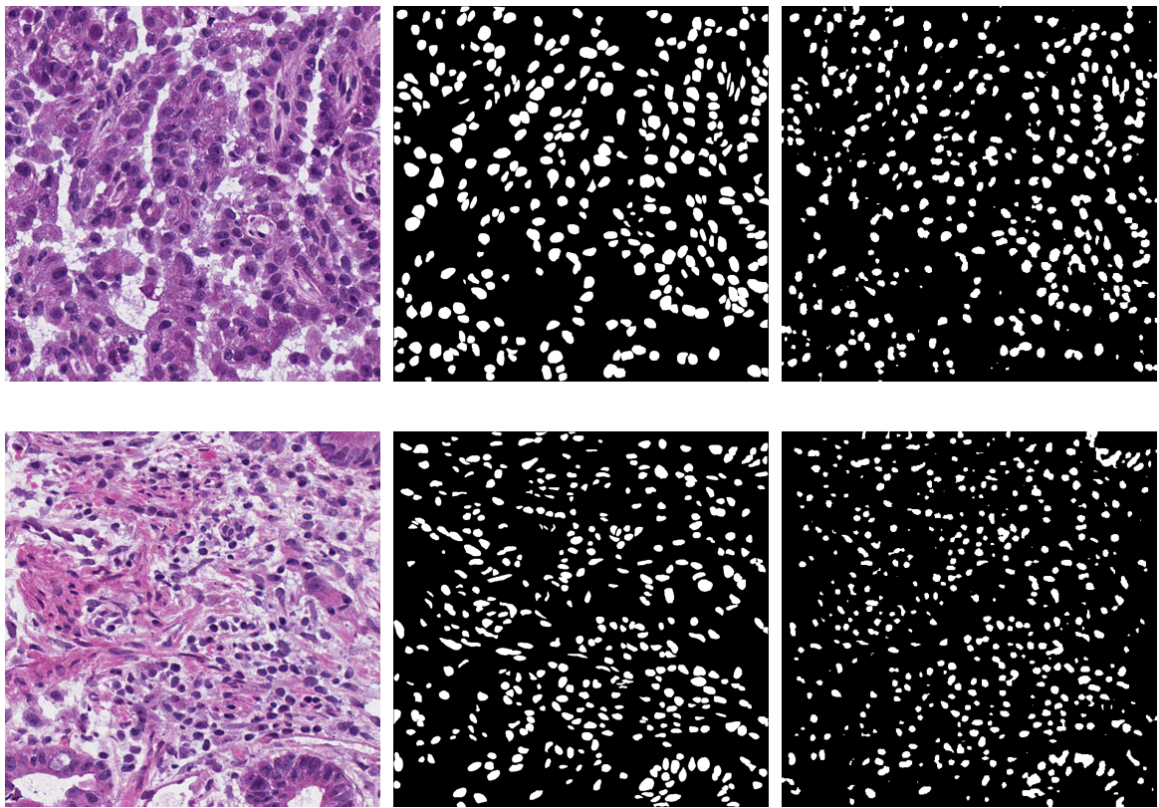


Fig. 4.16 *Semantic segmentation output for nuclei images. (Left) Original image. (Middle) Ground truth. (Right) Prediction of experiment (b) with DeepLab v3+ and backbone ResNet18.*

Nuclei Detection with Grad-CAM After the best-performing network has been identified, the output returned by the semantic segmentation was a mask in which the input image pixels were classified into pixels belonging to the foreground, i.e., nucleus, or background class. As mentioned previously, this did not allow the separation of multiple instances of the

same object. Therefore, the identification of multiple nuclei adjacent to each other was not possible.

In this scenario, the detection phase begins. In fact, after the semantic segmentation, post-processing was carried out in order to solve this problem. The first step was to calculate the Grad-CAM of the input image according to the chosen network. A CNN is often seen as a black box, or rather, as a model with parameters W that, given an image of input X , through a function $f(X, W)$, is able to map to the related output y . XAI techniques have been designed in order to unveil the underlying mechanisms involved in the processing stages of deep neural networks, and are recently gaining a lot of attention in medical imaging and clinical decision support systems [31, 130, 441, 442].

During the training phase, even if high performance, according to the considered metrics, was achieved, it was not known which image features are more determinant for the network to make its choices. One of the ways to visually solve this problem is Grad-CAM [31].

Grad-CAM is typically used in image-classification scenarios [132], but it can also be extended to semantic segmentation problems [443]. In general, the heatmap L^c for class c is generated by using a_c^k (as defined in Equation (4.6)) to sum the feature maps A^k , as in Equation (4.7).

$$a_c^k = \frac{1}{N} \sum_{u,v} \frac{\partial y^c}{\partial A_{uv}^k} \quad (4.6)$$

$$L^c = \text{ReLU} \left(\sum_k a_c^k A^k \right) \quad (4.7)$$

N is the number of pixels and (u, v) are the indices. ReLU is applied pixelwise to clip negative values at zero, to only highlight areas that positively contribute to the decision for class c . The difference with the classification task is that for semantic segmentation y^c , the scalar class score, is obtained by reducing the pixelwise class scores for the class of interest to a scalar [443], as in Equation (4.8).

$$y^c = \sum_{(u,v) \in P} Y_{(u,v)}^c \quad (4.8)$$

P is a set of pixel indices of interest in the output layer: in the case under examination, the softmax layer before the pixel classification layer. Higher values of L^c map denote which areas of the image are important for the decision to classify pixels.

In the proposed approach, the activation of the network for the nucleus class was analyzed, obtaining a probability map with values that will be denoted as CAM-Map. Therefore,

activations greater in correspondence with the centroids of the nuclei (even when adjacent to each other) are visible from Figure 4.17C.

From CAM-Map, a morphological grayscale dilation operator with a spherical shape factor of radius 7 was applied. The result is depicted in Figure 4.17D. This step allowed the enlargement of the activation areas so that no false nuclei were identified in the nearby regions where activations were not high enough compared to the maximum point.

Then, as portrayed in Figure 4.17E, it was carried out the calculation of the local maximum of the regions and the localization of all the connected components, with the related geometric centroids, which correspond to the identified nuclei.

Once the centroids were found, k -means clustering, with k equal to the number of connected components, was exploited to associate the adjacent pixels to each nucleus, so as to have the overall predicted mask of the original starting image. The final mask is reported in Figure 4.17F.

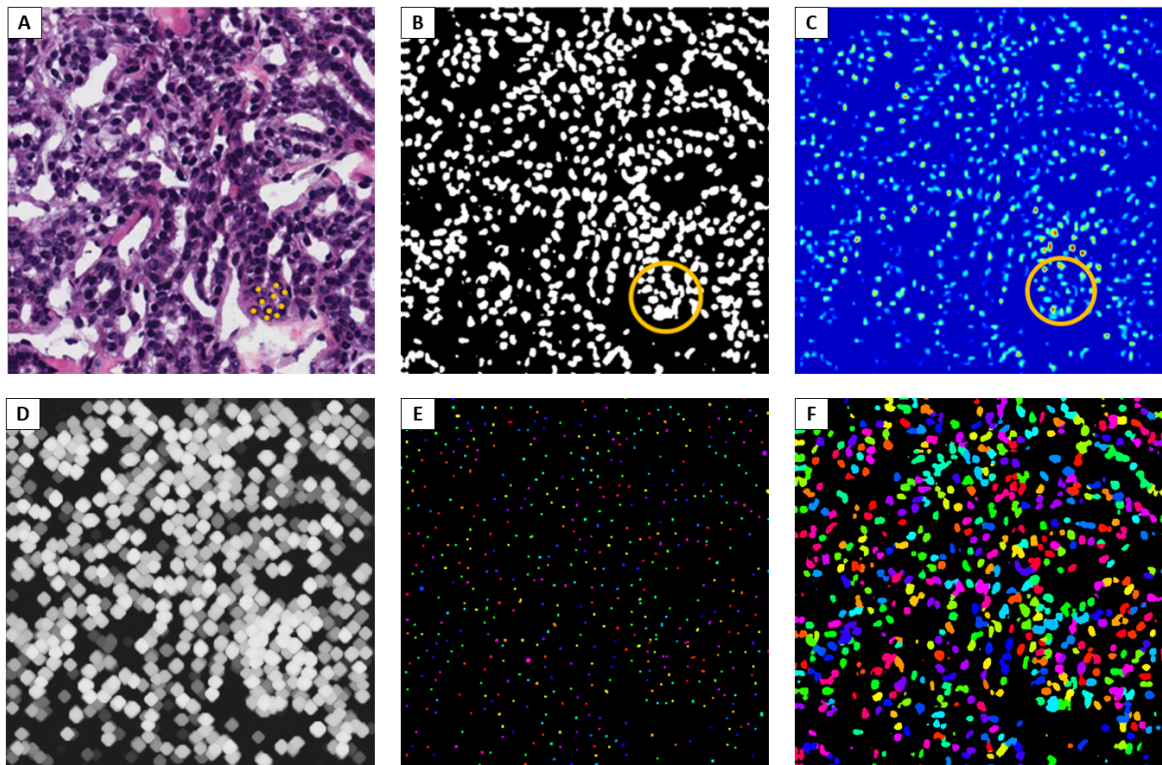


Fig. 4.17 *NDG-CAM Detection workflow*. (A) Zone with multiple neighboring instances of nuclei. (B) Failure to recognize adjacent nuclei. (C) Grad-CAM for semantic segmentation. (D) Dilated image. (E) Connected components. (F) Detection prediction.

4.2.1.3 Instance Segmentation

Object detection involves the detection, with a bounding box, of all the different objects of interest present in a scene. Instance segmentation further extends this task, by also considering the problem of delineating a precise mask around each object. Architectures for object detection are usually divided into one-stage and two-stage models, with the first being faster and the former being more accurate. Inside the realm of methods for two-stage object detectors, a pivotal role has been played by architectures from the R-CNN family [444].

Mask R-CNN evolves the R-CNN family by adding a semantic segmentation branch, making the model capable of performing instance segmentation [93]. The overall Mask R-CNN architecture is composed of two parts: the backbone architecture, which performs feature extraction, and the head architecture, which performs classification, bounding box regression, and mask prediction.

The Detectron2 [445] platform, powered by the Pytorch framework, has been adopted for the implementation of the Instance Segmentation model. Indeed, Detectron2 provides state-of-the-art detection and segmentation algorithms. It includes high-quality implementations of the most popular object detection algorithms, comprising different variants of the pioneering Mask R-CNN model. Detectron2 has an extensible design so that it can be easily employed to implement cutting-edge research projects.

The NuCLS dataset [439] was chosen to train the network, the instance segmentation model *mask_rcnn_R_50_DC5_1x*. Annotations were converted into the COCO annotation format for adoption in the Detectron2 framework.

4.2.1.4 Implementation Details

All the semantic segmentation networks have been trained on a laptop with a GeForce GTX960M. For carrying out the training, the chosen optimizer was SGDM, with a starting learning rate of 0.05. The learning rate schedule was piecewise with a drop factor of 0.94 and a drop period of 2. L_2 regularization parameter was set to 0.0005. With a batch size of 2, 15 epochs lasted roughly 105 min for the best-performing architecture, DeepLab v3+ with ResNet18 as the backbone.

The Mask R-CNN model, being heavier, has been trained on a Google Colab Pro environment. With a Tesla P100, 20,000 iterations were carried out in roughly 110 min. The chosen optimizer was SGDM, as set by default in the Detectron2 environment, with a starting learning rate of 0.00025.

4.2.1.5 Combined Model

In order to obtain the advantages of both approaches, a combined model has been developed.

It exploits a criterion for obtaining merged outputs from NDG-CAM detection and Mask R-CNN. In detail, a distance criterion was used to check if a nucleus was found by only one of the approaches. In that case, the nucleus was simply retained. Instead, if more nuclei centroids are found in proximity, only the ones found by Mask R-CNN are retained. The combined methodology has the idea to increase the recall, which is very important because nuclei detection is the first stage for further analyses.

4.2.1.6 Evaluation Metrics

Each semantic segmentation architecture previously described was tested in all three experimental configurations mentioned. In order to assess the goodness of pixelwise classification performed by semantic segmentation networks, the pixelwise precision, recall, and Dice coefficient were considered as performance indices.

Instead, for assessing the detection procedure, two kinds of metrics have been considered. The first is based on the simple calculation of the number of detected nuclei with respect to the ground truth. The error (e_a), defined in Equation (4.9), is given by the difference in absolute value between the number of nuclei found and the real number, divided by the latter. An example of the prediction vs. ground truth result, which is the basis for enumerating nuclei, is depicted in Figure 4.18A. To evaluate if the proposed algorithm was more prone toward over-detection or under-detection, a signed error (e_s), defined in Equation (4.10), was also evaluated:

$$e_a = \frac{|d - g|}{g} \quad (4.9)$$

$$e_s = \frac{d - g}{g}. \quad (4.10)$$

In these two equations, d denotes the number of detected nuclei, whereas g is the number of ground truth nuclei.

The second category of metrics includes the Dice coefficient, precision, and recall for object detection, which can provide more information about the quality of the detection results. In this case, metrics do not simply reward the count of predicted nuclei with respect to ground truth but are considering also the position.

In order to achieve this result, it was needed to discover object detection FP and FN, as can be seen in Figure 4.18B. In order to determine these quantities, as the first step, the distance matrix between the centroids of the detected nuclei and the real ones was computed.

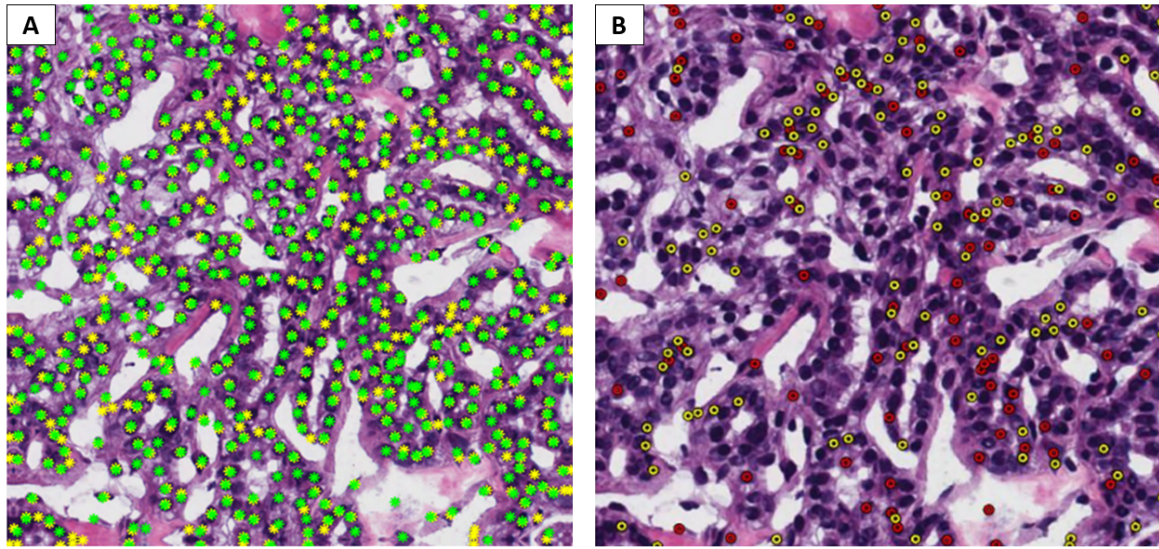


Fig. 4.18 *Example of calculation of evaluation metrics for object detection. (A) Prediction vs ground truth. Yellow, ground truth; green, prediction; (B) Differences between prediction and ground truth. Yellow, detection FN; red, detection FP.*

In order to decide whether a detection actually corresponds to a nucleus centroid, a distance threshold ξ was considered, equal to the mean radius of the nuclei of each image [434]. If the distance between a prediction and a ground truth annotation is less than or equal to ξ , the prediction is counted as a TP. If more than one detection verifies this condition, the one closest to the ground truth position is counted as TP and the others as FP. The detections further than ξ from any ground truth location are counted as FP, and all ground truth annotations without close detections are marked as FN. Lastly, the following control condition was added. If the distance between an FP and an FN is less than an ε threshold, set to 6 (a value close to the nuclear radius), the count of FP and FN will each be decreased by one, whereas TP will be increased by one. The pseudocode for determining TP, FP, and FN is reported in Algorithm 2.

In order to assess the statistical significance of the obtained results calculated per case, the p -value was determined with the two-tailed Wilcoxon signed-rank test. The threshold for significance has been set to 0.05.

4.2.2 Results

The automatic segmentation of cell nuclei attracted significant interest from the scientific community, as their identification is an important starting point for many medical analyses based on histopathological images. In this research work, for the semantic segmentation

Algorithm 2: Object Detection TP, FP, FN calculation.

```

input :  $gt$ , the ground truth nuclei centroids, an array of  $g$  coordinate pairs
          $pred$ , the predicted nuclei centroids, an array of  $d$  coordinate pairs
          $\xi$ , the mean radius of the ground truth nuclei
          $\varepsilon$ , the distance threshold // set to 6
output :  $TP$ , the true positives;  $FP$ , the false positives;  $FN$ , the false negatives
          $TP = FP = FN = 0; i = 0$ 
          $idx_{FP}, idx_{FN} = list(), list()$  // false positive/negative indexes
          $\delta = distance(gt, pred)$  // the distance matrix
while  $i < size(gt)$  do
     $v = \delta[:, i]$ 
     $idx = where(v < \xi)$  // a (possibly empty) array of indexes
    if  $size(idx) == 1$  then
         $TP = TP + 1$ 
    else if  $size(idx) > 1$  then
         $TP = TP + 1$ 
         $FP = FP + (size(idx) - 1)$ 
         $idx_{FP}.extend(idx)$ 
    else if  $size(idx) == 0$  then
         $FN = FN + 1$ 
         $idx_{FN}.append(i)$ 
    end
     $i = i + 1$ 
end
 $arr_{FN} = filter(gt, idx_{FN})$  // extract the false negatives
 $p = 0$ 
while  $p < size(idx_{FP})$  do
     $a = 0$ 
    while  $a < size(arr_{FN})$  do
         $\Delta = distance(pred[p], arr_{FN}[a])$ 
        if  $(\Delta \leq \varepsilon)$  then
             $FP = FP - 1$ 
             $FN = FN - 1$ 
             $TP = TP + 1$ 
        end
         $a = a + 1$ 
    end
     $p = p + 1$ 
end

```

phase, different architectures were elaborated and tested on different datasets, for a total of 15 experiments. For each of them, performance indices were calculated to identify the best

model with which to proceed for the subsequent phases. From this comparison, it emerged that the best performance can be obtained by referring to the experimental configuration (b) defined in Section 4.2.1.2.

Table 4.7 reports the results obtained for each network architecture in the semantic segmentation task. For DeepLab v3+, the backbone architecture is included within square brackets.

Table 4.7 *Performance comparison between considered network architectures for semantic segmentation.*

Network	Metric	Experiment (a)	Experiment (b)	Experiment (c)
U-Net	DICE	66.74 ± 3.44	65.71 ± 8.57	60.74 ± 11.65
	PRECISION	57.13 ± 8.15	52.69 ± 11.96	45.43 ± 11.77
	RECALL	83.56 ± 10.61	91.65 ± 6.57	96.46 ± 2.44
SegNet	DICE	56.44 ± 9.31	65.05 ± 6.32	62.02 ± 12.28
	PRECISION	67.09 ± 8.01	58.93 ± 14.23	51.67 ± 14.96
	RECALL	52.60 ± 16.20	81.35 ± 17.69	85.05 ± 13.24
DeepLab v3+ [ResNet18]	DICE	52.21 ± 11.99	74.23 ± 4.85	72.17 ± 8.03
	PRECISION	76.78 ± 6.60	76.42 ± 8.69	62.76 ± 11.78
	RECALL	41.76 ± 13.55	74.25 ± 11.23	87.17 ± 5.64
DeepLab v3+ [ResNet50]	DICE	57.87 ± 6.88	61.68 ± 8.75	65.98 ± 7.84
	PRECISION	59.70 ± 6.35	63.69 ± 7.51	54.14 ± 13.81
	RECALL	57.10 ± 10.43	60.71 ± 11.94	90.95 ± 10.02
DeepLab v3+ [mobilenetv2]	DICE	56.64 ± 6.60	73.01 ± 7.56	66.31 ± 13.80
	PRECISION	66.49 ± 5.56	73.50 ± 11.76	57.52 ± 16.31
	RECALL	50.66 ± 10.50	75.07 ± 10.38	85.35 ± 9.43

Therefore, it emerges that the best solution coincides with experiment (b) conducted with DeepLab v3+ using the ResNet18 network as the backbone. It allowed the attainment of a pixelwise Dice coefficient of $74.23 \pm 4.85\%$, a precision of $76.42 \pm 8.69\%$, and a recall of $74.25 \pm 11.23\%$.

DeepLab v3+ was hence chosen as the base model to be exploited in the detection phase. By exploiting the Grad-CAM for semantic segmentation, it was possible to retrieve nuclei centroids via local maxima of the obtained saliency maps.

On the V1 dataset, the experimental results demonstrated an e_a of the identified nuclei equal to 2.11%, 2.43%, and 11.50% for the NDG-CAM, Mask R-CNN, and combined method, respectively. When calculated per case, the values for e_s were $1.84 \pm 13.05\%$, $3.46 \pm 6.15\%$, and $14.45 \pm 11.22\%$, indicating that the models generally tend to over-detect on this dataset.

In the V4 dataset, the e_a had a value of 15.26%, 59.22%, and 14.10% for the NDG-CAM, Mask R-CNN, and combined method, respectively. When calculated per case, the values for e_s were $-16.86 \pm 13.79\%$, $-60.13 \pm 13.88\%$, and $-14.88 \pm 12.86\%$, showing that the models have a tendency to under-detect on this dataset. In particular, it was noticed that very small nuclei, such as those of lymphocytes, and elongated ones, such as those of fibrocytes, were under-detected.

For the detection task, the results are reported in Table 4.8. In the V1 dataset, NDG-CAM, Mask R-CNN, and the combined method were capable of achieving a Dice coefficient of 0.824, 0.878, and 0.884, respectively. Thus, the combined method obtained slightly better results than the other methods. As for the recall, the combined method decisively surpasses the other approaches, with a value of 0.934.

In the V4 dataset, the combined method proves to be the best, achieving a recall of 0.850 and a Dice coefficient of 0.914. Mask R-CNN performs poorly in this case, with a recall of 0.403 and a Dice coefficient of 0.573.

Table 4.8 *Comparison of methods for Nuclei Detection*. This extends the one proposed by Alom et al. [340] and Sirinukunwattana et al. [347].

Method	Precision	Recall	Dice
CRImage [345]	0.657	0.461	0.542
CNN [345]	0.783	0.804	0.793
SSAE [342]	0.617	0.644	0.630
LIPSyM [346]	0.725	0.517	0.604
SC-CNN (M = 1) [347]	0.758	0.827	0.791
SC-CNN (M = 2) [347]	0.781	0.823	0.802
UD-Net [340]	0.822	0.842	0.828
NDG-CAM (V1)	0.833	0.815	0.824
NDG-CAM (V4)	0.992	0.841	0.910
Mask R-CNN (V1)	0.867	0.888	0.878
Mask R-CNN (V4)	0.989	0.403	0.573
Combined (V1)	0.838	0.934	0.884
Combined (V4)	0.986	0.850	0.914

The violin plots calculated per tile are reported in Figure 4.19 for the V1 and V4 datasets, comparing the NDG-CAM detection method, Mask R-CNN, and the combined approach. It is worth noting that the Mask R-CNN model works very well on the V1 dataset but performs poorly on the V4 one. On the other hand, the NDG-CAM and the combined methods maintain high levels of performance in all scenarios.

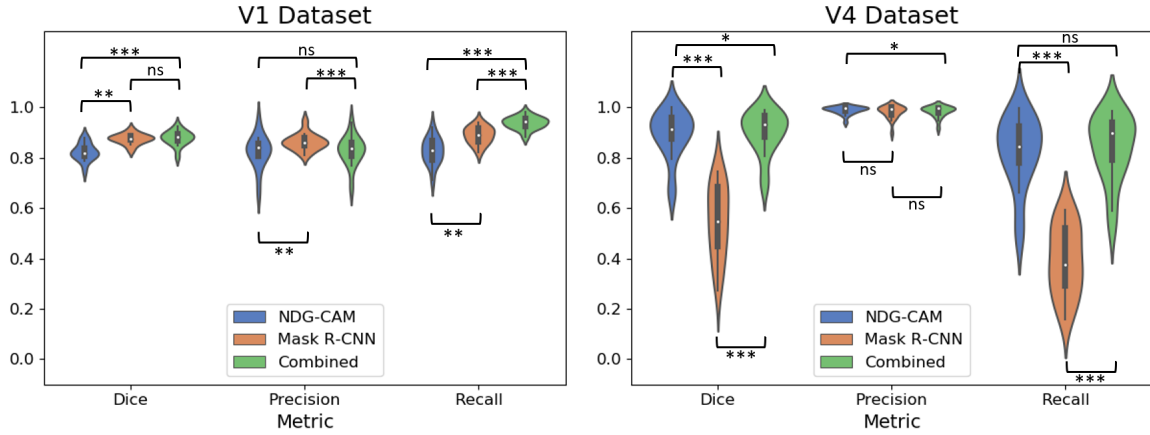


Fig. 4.19 Violin plots for the detection metrics calculated per case. (Left) V1 dataset. (Right) V4 dataset. In the figure, ns stands for nonsignificant; * denotes p -value < 0.05 ; ** indicates p -value < 0.01 ; and *** means p -value < 0.001 .

In the V1 dataset, the combined model does not show a Dice coefficient that is higher in a statistically significant way than the Mask R-CNN approach, with a p -value of 0.07. On the other hand, the recall was much higher for the combined method, resulting in a p -value < 0.001 for both NDG-CAM and Mask R-CNN. In the V4 dataset, both the NDG-CAM and the combined method showed much stronger results than Mask R-CNN, with a p -value less than 0.001 in both cases for Dice coefficient and recall. Moreover, the combined approach shows a statistically significant advantage over NDG-CAM (p -value = 0.048) for the Dice coefficient.

4.2.2.1 Discussion

In order to show the effectiveness of the proposed method, it was compared with existing state-of-the-art approaches. It has to be noted that the presented method allows exploiting semantic segmentation architectures to realize nuclei detection, whereas other approaches usually involve networks specialized for this task. Several approaches proposed in the literature try to localize centers of the nuclei or proximity maps to those centers [23, 347, 434]. These approaches require instance-level annotations, although the results are promising. On

the other hand, the proposed method exploits an XAI technique, Grad-CAM for semantic segmentation, to reconstruct post hoc saliency maps that are related to the centers of the nuclei, showing that semantic segmentation networks can perform detection tasks without specialized modifications.

The most widespread metrics employed for assessing algorithms for object detection involve precision, recall, and the Dice coefficient. Namely, they are the metrics that are also related to the position of the detected nuclei, and not only to the counts.

A quantitative comparison between considered approaches and existing ones from the literature is presented in Table 4.8.

From this comparative analysis, it emerges that the proposed method is perfectly aligned with the state-of-the-art, without the need to implement specific kinds of specialized loss functions [72] or architectures for detection [19, 93].

Indeed, the NDG-CAM method alone was capable of achieving a Dice coefficient for object detection of 0.824, whereas the UD-Net [340] method, the top-performing method among the selected from the literature, had a Dice coefficient of 0.828. When the proposed NDG-CAM detection method is used in combined usage with Mask R-CNN, the recall increases to 0.934, and the Dice coefficient to 0.884, surpassing the current state-of-the-art methods for nuclei detection. On the collected external validation set, metrics are even higher, with a Dice coefficient of 0.914, showing the generalization capabilities of the proposed workflow.

Qualitative results for the object detection pipeline involving semantic segmentation and Grad-CAM on the images of the independent external validation set V4 are depicted in Figure 4.20. Instead, Figure 4.21 shows the final detection results on the validation datasets V1 and V4 with the NDG-CAM method, the Mask R-CNN architecture, and the combined adoption of both methods.

It can be seen from the images of Figure 4.20, taken from the V4 dataset, that precision is very high. Indeed, virtually all detected nuclei are real. Some small or elongated nuclei, such as lymphocytic or fibrocytic nuclei, are under-detected. This may be due to a lack of proper training datasets with a large variety of nuclear shapes.

The two methods show similar performance on the V1 dataset, as can be observed from Figure 4.21. Mask R-CNN achieves slightly better performance on this dataset, and considering that it has been trained on a larger training set, the combined method proved to be superior. From the same figure, it is possible to observe that, in the V4 dataset, Mask R-CNN does not properly generalize, resulting in the missing of many nuclei (low recall).

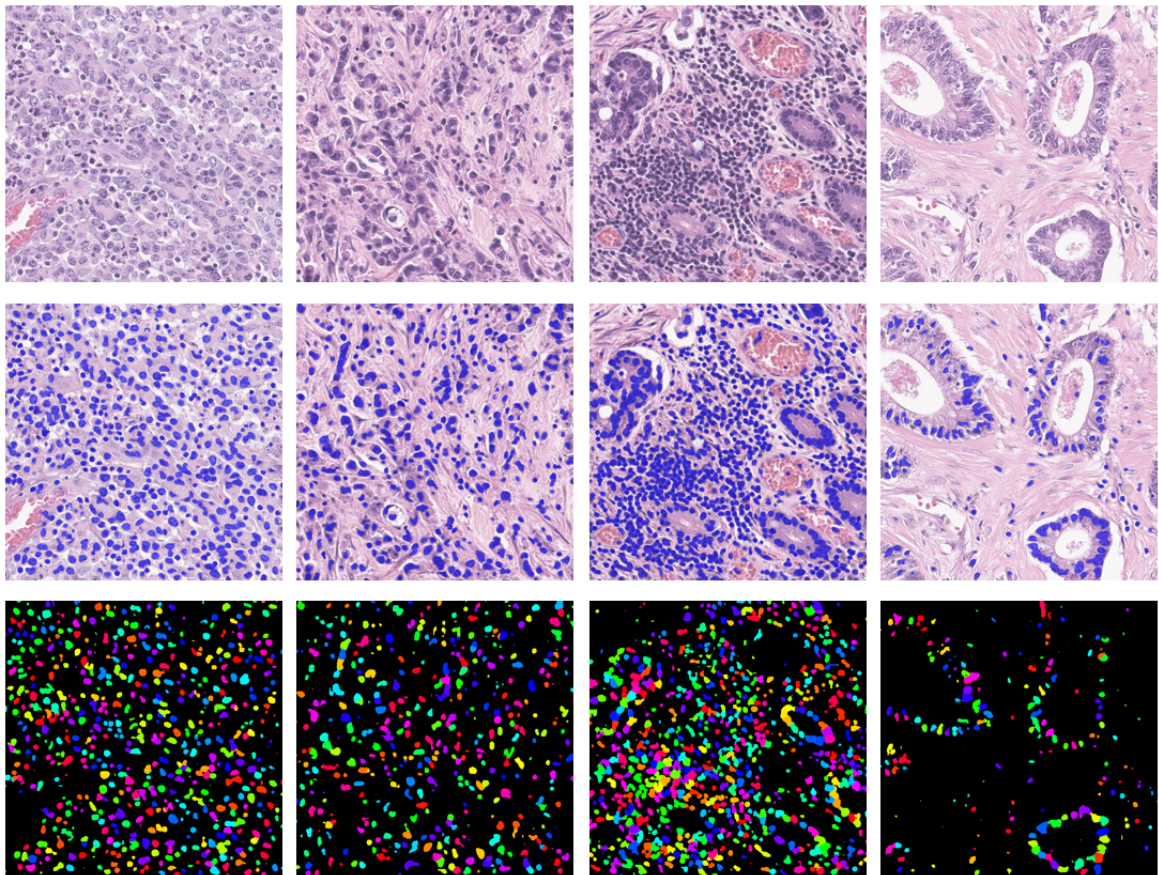


Fig. 4.20 *Examples of the NDG-CAM method on the data from the Pathology Department of IRCCS Istituto Tumori “Giovanni Paolo II”.* Results are shown for the best architecture (DeepLab v3+ with ResNet18 backbone). (**First row**) Original images. (**Second row**) Semantic segmentation. (**Third row**) Instance segmentation after detection of centroids of the nuclei, with each color denoting a different nuclear instance.

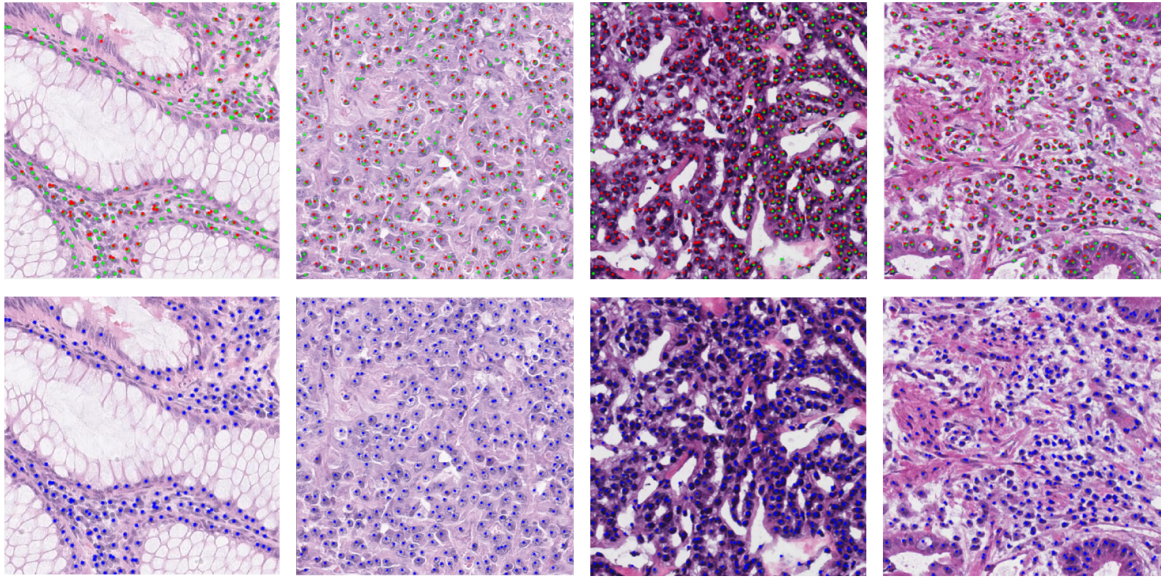


Fig. 4.21 *Examples of centroid detection on the validation sets V1 and V4. (Top row) Green, NDG-CAM method detections; red, Mask R-CNN detections. (Bottom row) Blue, combined method detections. The first and second columns show data from V4, whereas the third and fourth columns depict data from V1.*

4.2.3 Summary

In the research work described in this section, a novel method was presented with the aim of nuclei identification from histological H&E images. In the proposed multi-stage pipeline, the first phase involved semantic segmentation. After various experiments, DeepLab v3+ (ResNet18 backbone) emerged as the best-performing architecture. Subsequently, because this analysis did not allow the distinction of multiple instances of the same object, a novel detection algorithm, NDG-CAM, which exploited Grad-CAM to solve the problem of separating the instances, has been proposed. Even without the need to use specialized loss functions or architectures, it allowed to achieve satisfactory results in the detection task, comparable to or even better than more sophisticated training setups [23, 342, 345, 434]. When the method is combined with the Mask R-CNN instance segmentation architecture, results exceed the state-of-the-art methods for nuclei detection.

Even though the local validation set includes only colorectal cancer H&E slides, it has to be considered that in each slide there are several tissue types present (e.g., stroma, immune infiltration) and the proposed method has the ability to detect nuclei not only related to the tumor or normal epithelium of colon but also to other cytotypes.

Indeed, it was noticed under-detection of lymphocytic or fibrocytic nuclei, and this could be explained by a lack of datasets enriched in these nuclei subtypes. For such a reason, a

direction for future works includes the collection of a dataset with multiple and balanced nuclei annotations.

On the clinical side, the proposed workflow could be a valid tool to support pathologists in the detection and reporting of histological samples, thus allowing a considerable saving of time and resources, besides providing an objective tool that is more reliable than manual assessment. Future works will concern the classification of the detected nuclei, in order to estimate how many are malignant or subjected to specific lesions, so that important clinical parameters, such as neoplastic cellularity, can be determined quantitatively.

4.3 Computer-Aided Diagnosis for IgA-GN Oxford Classification

In this section, a novel CAD system, to predict the status of four compartments of the Oxford classification (M-score, E-score, S-score, and C-score), is proposed. The definition of the MEST-C-score can be found in Section 2.3.3. Two approaches are considered: the first is based on image processing filters and feature extraction procedures, followed by layers that carry out statistical classification of the data, whereas the second exploits state-of-the-art DL classifiers.

The existing literature based on supervised approaches, surveyed in Section 2.3.3, focused on a few aspects of the MEST-C-score. Purwar et al. [351] considered only M-score, Chagas et al. [352] dealt with E-score and M-score, whereas Weis et al. [353] analyzed S-score and M-score. An unsupervised approach has been tried by Sato et al. [355], showing interesting clustering results. Uchino et al. [354] tackled the task of scoring the different compartments of the MEST-C-score, however, the approach proposed in this section managed to achieve better performance.

The remainder of this section is structured as follows: Section 4.3.1 introduces the dataset exploited for the experiments. Section 4.3.2 describes the techniques adopted for carrying out the analysis, both the image processing and machine learning-based, and the DL ones. Results are presented in Section 4.3.3, whereas conclusions and final remarks are drawn in Section 4.3.4.

4.3.1 Materials

To enhance the contrast in samples examined under the microscope, it is essential to choose an adequate staining technique. The most common staining techniques include H&E, PAS, Periodic acid-methenamine silver (PAM), and so on.

Furthermore, sometimes more stainings are applied to obtain complementary information. For a routine diagnosis the typical choice is H&E because it gives the first impression on the tissue composition (cortical and medullary, cell infiltrations, number of glomeruli, and so on). The PAS staining is the most sensible choice for a more detailed analysis of the glomerulus since it highlights glomerular cells and potential modifications of glomerular structure. In this study, all images are PAS stained since it is the preferred staining technique by pathologists for analyzing glomerular structures.

4.3.1.1 Dataset Composition

The dataset used in the study comes from the Institute of General Pathology and Pathological Anatomy of Uniklinik Köln in Cologne (Germany).

During the analysis phase, the following scores were considered: (i) Score M, coded as "Indeterminate" ("NULL") / "noM" / "yesM", depending on whether the ratio of mesangial cells / mesangial area is indeterminate, < 4 or ≥ 4 ; (ii) Score E, represented as "noE" / "yesE", depending on the absence or presence of the condition of endocapillary hypercellularity; (iii) Score S, referred to "NoGS" / "SGS" / "GGS", depending upon the absence of glomerulosclerosis or the presence of segmental and global glomerulosclerosis, respectively; (iv) Score C, illustrated as "noC" / "yesC", depending on the absence or presence of the cellular or fibrocellular crescent condition. Please note that the definition of segmental glomerulosclerosis employed for the collection of this dataset is different from the usual definition. Glomeruli with any amount but not the whole of the tuft involved in sclerosis were labeled as "SGS".

The original dataset consists of 3617 images of ROIs centered around glomeruli, with the related segmentation masks, so that further stages can be focused only on the glomeruli regions. Due to the impossibility of properly labeling some of the images with glomeruli for problems like artifacts or incomplete ROI, the final dataset on which the processing and the consequent classifications are carried out, turned out to be 3575 images with the related masks. The distribution of the considered classes in the dataset is depicted in Figure 4.22, whereas sample images from the dataset are reported in Figure 4.23.

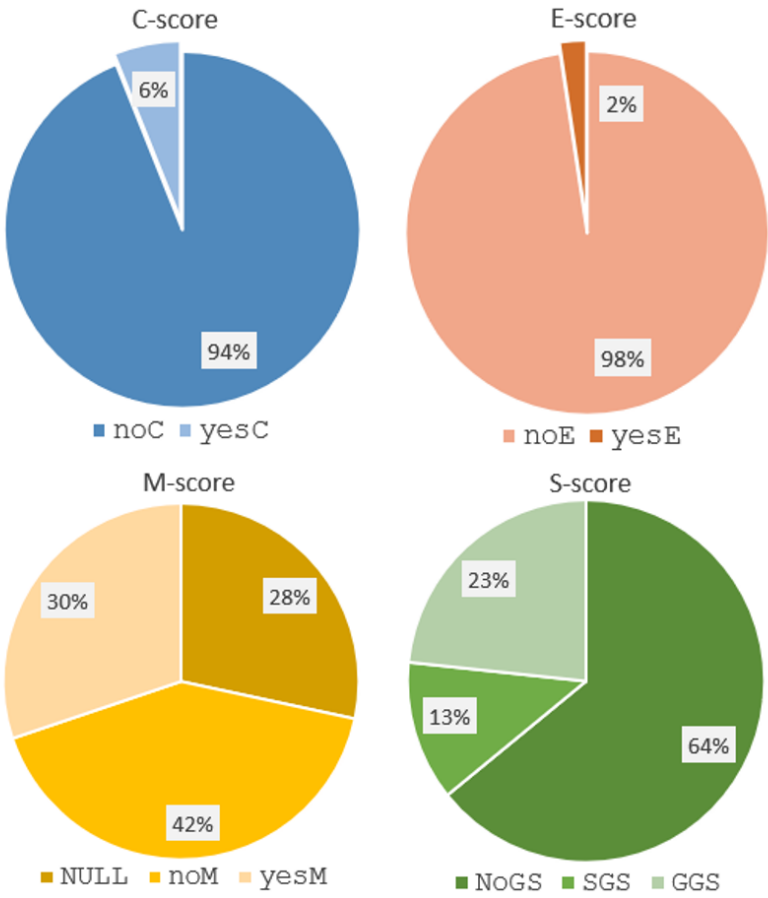


Fig. 4.22 The glomeruli dataset distribution.

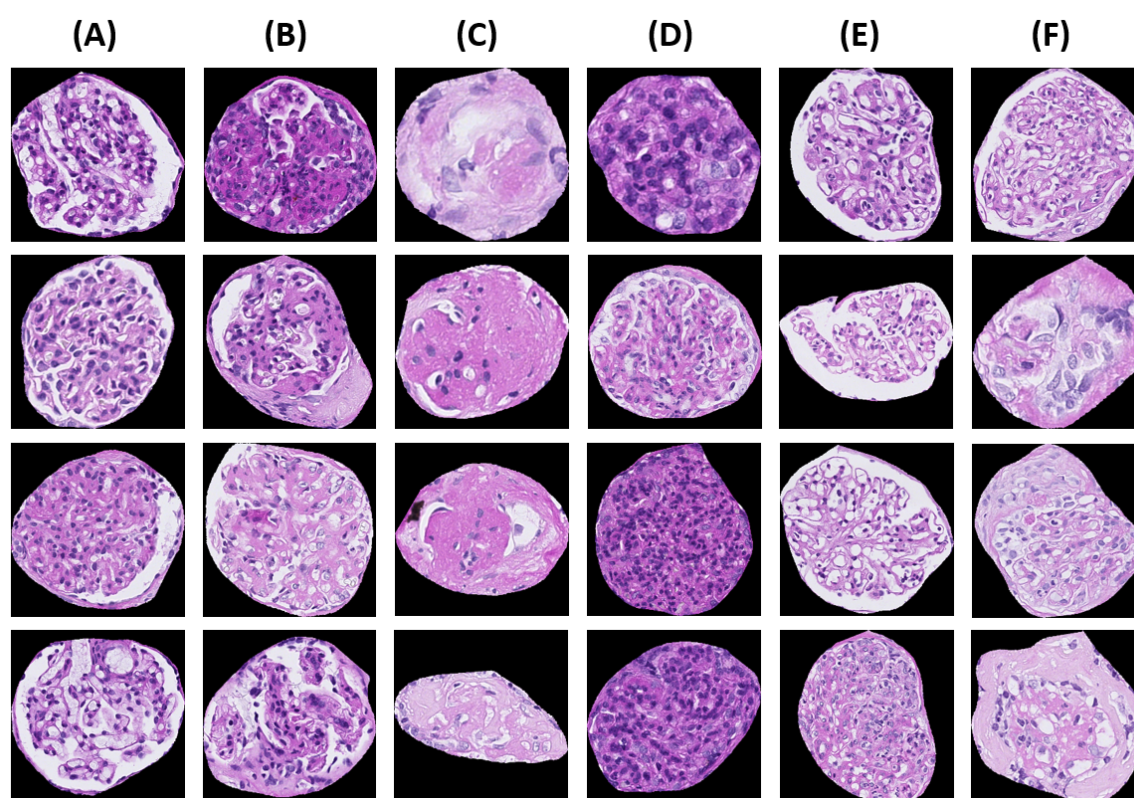


Fig. 4.23 *IgA Glomerulonephritis — Glomerular patterns.* (A) Healthy Glomeruli; (B) Segmental Glomerulosclerosis; (C) Global Glomerulosclerosis; (D) Mesangioproliferation; (E) Endocapillary Hypercellularity; (F) Cellular or Fibrocellular Crescent.

4.3.2 Methods

To perform the classification of the different lesions considered for the glomeruli, two different approaches have been employed, the first one based on classical feature extraction, and the second one based on deep features extracted from CNNs. Lastly, ANN models have been trained on those features to perform the classification. In both cases, the classification related to the C-score and the E-score has been treated as a binary classification problem, whereas the classification related to the M-score and the S-score has been treated as a three-classes classification problem, according to the convention introduced in Section 4.3.1.1.

4.3.2.1 Classical Features

The considered classical feature set includes morphological, textural, and histogram features extracted from the images. Then, those features have been used to train ANNs. The image processing procedures were performed with C++ using the OpenCV library. The pre-processing and feature extraction workflow is shown in Figure 4.24.

In the preprocessing stage, the first operation consisted of applying the morphological opening to the masks, in order to remove eventual small noisy points outside the glomerulus region. Thereafter, the images have been masked, cropped, padded to a square size, and rescaled to 400×400 .

Color channels extraction has been performed, both in RGB and CMYK color spaces. Since green (RGB) and magenta (CMYK) image channels have proven to be more representative of the glomerular structure, all the subsequent operations have been applied to these two channels. For color images with 8-bit per channel, CMYK channels can be computed from the RGB ones using the following relations: $R' = \frac{R}{255}$; $G' = \frac{G}{255}$; $B' = \frac{B}{255}$; $K' = \min(1 - R, 1 - G, 1 - B)$; $C = \frac{1-R-K}{1-K} \times 255$; $M = \frac{1-G-K}{1-K} \times 255$; $Y = \frac{1-B-K}{1-K} \times 255$.

The images have been equalized using both the traditional *histogram equalization* and the *contrast limited adaptive histogram equalization* (CLAHE) [446]. The advantage of employing Adaptive Histogram Equalization (AHE) consists of exploiting different histograms for different regions of the image, so that locality is considered for the redistribution of lightness values. On the other side, AHE has an important drawback: it can worsen noise in homogeneous regions of the image. In order to overcome this limitation, a variant of AHE has been proposed, CLAHE, which limits the amplification by clamping the histogram to a given value, the clip limit, before calculating the cumulative distribution function. The clip limit is a parameter that is calibrated according to the size of the neighborhood region.

To perform these techniques the OpenCV library has been used, in particular the function *equalizeHist()* and the *CLAHE* class. The clip limit has been set to 4. The images have been partitioned into squared tiles of 8×8 pixels.

The *Laplacian of Gaussian* operator (LoG) has been implemented via the OpenCV *Laplacian()* and *GaussianBlur()* functions, where the Gaussian filter has been chosen with a 3×3 kernel.

The *Discrete Wavelet Transform* (DWT) is a pivotal method for signal processing and analysis. Its efficiency has been proven for numerous applications, such as image denoising and compression. The properties which make DWT important include (i) the capability to analyze signals with time-varying characteristics, since it allows to perform a fine time-frequency localization; (ii) the possibility to realize multi-resolution representations; (iii) simplicity of implementation through a filter bank.

The DWT has been computed through the iterative implementation of two high-pass and low-pass filters: $HPF : \frac{1}{\sqrt{2}}[1, -1]$; $LPF : \frac{1}{\sqrt{2}}[1, 1]$

The output consists of an image divided into four sub-bands:

- The LL sub-band is the approximation band and is obtained by applying the convolution with the low-pass filter in the horizontal and vertical directions.
- The LH sub-band (the first of the three detail sub-bands) is the output of the convolution with a low-pass filter in the horizontal direction and with a high-pass filter in the vertical direction.
- The HL sub-band (the second of the three detail sub-bands) is the output of the convolution with a high-pass filter in the horizontal direction and with a low-pass filter in the vertical direction.
- Finally, the HH sub-band (the last detail sub-band) is the output of the convolution with a high pass filter both in the horizontal and vertical directions.

The purpose of applying different equalization and filtering methods to the images was to find the techniques leading to the most discriminative features.

Hu moments and *Haralick texture features* have been extracted from different sources: images equalized with the classical histogram equalization, images equalized using CLAHE, images filtered with LoG applied after CLAHE, all the four sub-bands derived from the application of the wavelet transform to the images formerly equalized using CLAHE.

Seven *histogram features* (mean, variance, standard deviation, skewness, kurtosis, entropy, and energy) have been extracted from images filtered with LoG applied after CLAHE and

all of the four sub-bands derived from the application of the discrete wavelet transform to the images formerly equalized using CLAHE. All these operations have been computed considering both magenta and green color channels.

In addition, four *morphological features* (compactness, eccentricity, roundness, and convexity) have been extracted from the masks. Let a be the area, p the perimeter, l the length of the short axis, L the length of the long axis, c the convex perimeter, then those features can be defined as in Equation (4.11), Equation (4.12), Equation (4.13), and Equation (4.14), respectively.

$$Compactness = \frac{4 \cdot \pi \cdot a}{p^2} \quad (4.11)$$

$$Eccentricity = \frac{l}{L} \quad (4.12)$$

$$Roundness = \frac{4 \cdot \pi \cdot a}{c^2} \quad (4.13)$$

$$Convexity = \frac{c}{p} \quad (4.14)$$

All the ANNs have been developed using MATLAB[®] with slightly different topology architectures depending on the E-score under consideration.

For the C-score and the E-score an ANN with the following options has been built with *hiddenLayerSize* set to `[10 10]`, *trainingFunction* set to `trainidx`, *epochs* set to 500, *costFunction* set to `crossentropy`, and *lastLayerActivationFunction* set to `logsig`.

The data were split into a training set (80%) and a testing set (20%). Furthermore, 20% of data from the training set has been used for validation. Due to its underrepresentation, the positive class has been oversampled in both cases using the ADASYN [447] method. The oversampling has been computed just for the training set. The different tests were carried out by extracting three kinds of features: Hu moments (Hm), Haralick features (Hf), and histogram features (Hi). To avoid overfitting, the *lambda* regularization parameter has been set as follows: (i) $\lambda = 0.9$ when the features used were Hm and Hf extracted from the equalized images only; (ii) $\lambda = 0.6$ when the features used were Hm and Hf extracted from the equalized images (CLAHE) together with the Hi computed on the LoG applied to the equalized images; (iii) $\lambda = 0.5$ when the features used were Hm and Hf extracted from the equalized images (CLAHE) together with the Hi computed on each of the four sub-bands deriving from the DWT application.

For the M-score and the S-score an ANN with the following options has been built: *hiddenLayerSize* set to [20 20], Adam [256] optimizer, batch size of 64, and *GradientThresholdMethod* set to *l2norm*. The last layer is a softmax layer because in these cases.

The following features' configurations were compared:

- Hm and Hf extracted from the images equalized with the traditional histogram equalization (*histeq()* in MATLAB[®]);
- Hm and Hf extracted from the images equalized using CLAHE;
- Hm and Hf calculated on the LoG applied to the image equalized with CLAHE;
- Hm and Hf calculated on each sub-band deriving from the application of the discrete wavelet transform to the image equalized with CLAHE;
- Hm and Hf calculated on the image equalized with CLAHE and Hi calculated on the LoG applied to the image equalized with CLAHE;
- Hm and Hf calculated on the image equalized with CLAHE and Hi calculated on each DWT sub-band of the image equalized with CLAHE.

In all cases, the morphological features calculated on the masks were also included.

The best results were obtained in all cases by calculating the Hu moments and the Haralick features on the image equalized with CLAHE and the features of the histogram calculated on LoG (in the case of C-score and E-score) and on the sub-bands of the DWT (in the case of M-score and S-score).

To compare the performance of the ANNs, in order to find the image processing techniques leading to the most discriminative features, the following metrics have been considered: accuracy, sensitivity, and precision.

4.3.2.2 Convolutional Neural Networks

With regard to deep learning and convolutional neural networks, the first approach taken into account for this study has been *transfer learning*.

In transfer learning [57], a model previously trained to carry out a specific task is reused for the execution of a second different job. The idea is to exploit the data of the first iteration of the neural model training to extract information that can be useful for making new predictions on data from a different domain. The intuition behind transfer learning, especially in the image classification domain, is that if a model is trained on a sufficiently

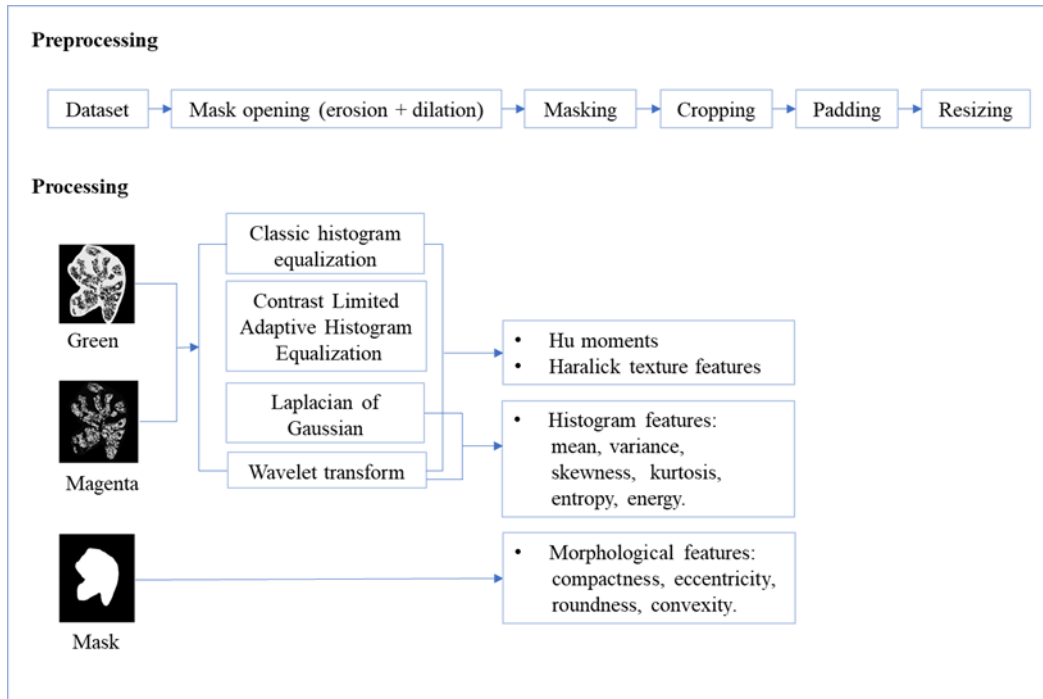


Fig. 4.24 Workflow for preprocessing and feature extraction from images.

large and general dataset, as ImageNet [190], it will effectively be capable to extract universal representations of the visual world; it will therefore be possible to exploit the feature maps learned without having to train a new model from scratch, wasting resources and time in the training process. The employed transfer learning workflow is reported in Figure 4.25.

Features have been extracted from the last convolutional layer of the following pre-trained CNNs: ResNet-18, ResNet-50 [48], GoogleNet [43], SqueezeNet [66], and MobileNet [448] for C-score, E-score and M-score. Moreover, for S-score also ResNet-101 [48], VGG-16 [41], AlexNet [26], and Inception-v3 [64] have been considered.

The feature matrices extracted from each of the aforementioned pre-trained models have been used to train an ANN with a different architecture depending upon both the target lesion and the considered model. The different topologies exploited for the ANNs are reported in Table 4.9, considering all CNN architectures and endpoints.

The following options have been also set for all cases regarding C-score and E-score: *trainingFunction* set to *trainidx*, 500 epochs, *costFunction* set to *crossentropy*, and *lastLayerActivationFunction* set to *logsig*.

As for the classical features models, the data were split into a training set (80%) and a testing set (20%). Furthermore, 20% of data from the training set has been used for validation. Due to its underrepresentation, the positive class has been again oversampled in both cases

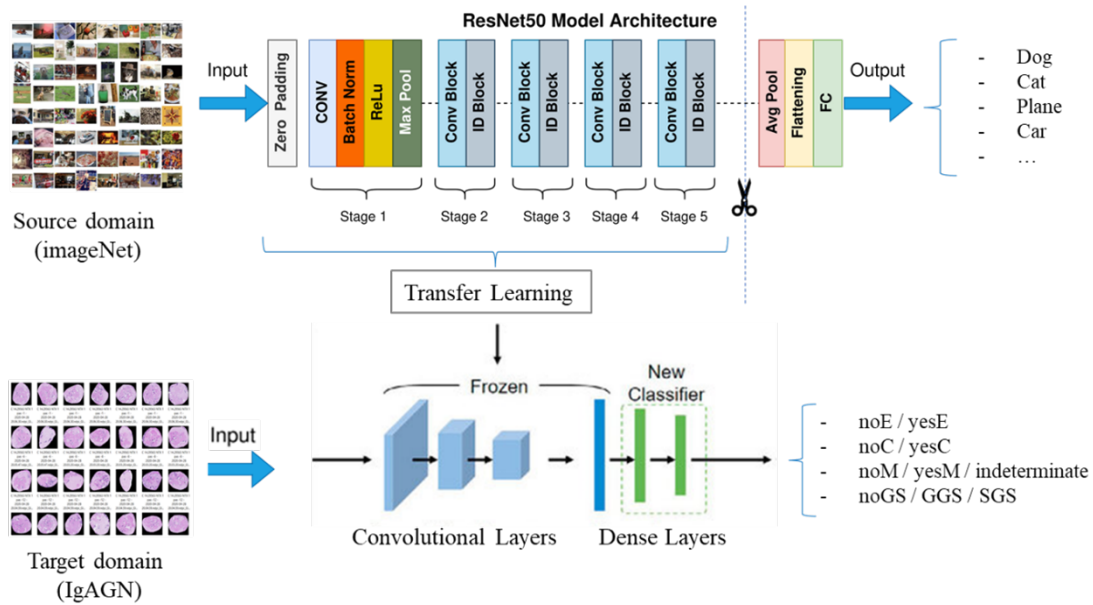


Fig. 4.25 Transfer Learning Workflow for Glomeruli MESC Classification.

Table 4.9 Topologies for the ANN configuration. Considered endpoints: M, E, S, C lesions.

Model	Score M	Score E	Score S	Score C
ResNet18	[64 32 8]	[15]	[64 32 8]	[15]
ResNet50	[64 32 8]	[15]	[64 32 8]	[10 10]
GoogLeNet	[64 32 8]	[15 10 5]	[64 32 8]	[15]
ResNet101	X	X	[128 32 64]	X
VGG16	X	X	[64 32 8]	X
AlexNet	X	X	[64 32 8]	X
InceptionV3	X	X	[64 32 8]	X
SqueezeNet	[64 32 8]	[10 10 10]	[64 32 8]	[10 10 10]
MobileNet	[64 32 8]	[10 10]	[128 32 64]	[15]

using the ADASYN method. The oversampling has been computed just for the training set. To avoid overfitting regularization has been taken into account.

The following options have been also set for all cases regarding M-score and S-score: *trainingFunction* set to *trainscg* and *net.layers{end}.transferFcn* set to *softmax*. As already mentioned, M-score and S-score classification have been treated as a three-classes classification problem.

The performances of the different approaches using pre-trained CNNs, the custom CNN and retrained ResNet-50 have been evaluated with the following metrics: accuracy, sensitivity, specificity, precision, F_1 -score, and AUC.

4.3.3 Results

4.3.3.1 Classical Features

Comparing the values of accuracy, sensitivity (recall) and precision, the features which have been proven to be especially discriminative are the concatenation of the following features extracted for both the green and the magenta channels:

1. Hm and Hf computed on images equalized using CLAHE;
2. Hi extracted from the sub-bands derived from the DWT applied to the equalized images (for C-score and E-score) or Hi extracted from the LoG applied to the equalized images (for M-score and S-score); Morphological features extracted from the masks have been also concatenated to the above-mentioned ones.

Numerical results, namely accuracy, precision, and recall, are reported in Table 4.10, Table 4.11, Table 4.12, and Table 4.13 for M-score, E-score, S-score, and C-score, respectively.

4.3.3.2 Convolutional Neural Networks

Numeric results, in terms of Accuracy, Precision, Sensitivity, Specificity, F_1 -score, and AUC are reported for all the transfer learning CNN architectures considered, for all the endpoints. Table 4.14, Table 4.15, Table 4.16, and Table 4.17 displays the results with CNN trained on M-score, E-score, S-score, and C-score.

The retrained ResNet-50 model has been exploited also to understand the regions behind the decisions made by the CNN, by exploiting the popular Grad-CAM [132] algorithm. Results of Grad-CAM are reported in Figure 4.26.

Table 4.10 *M-score — ANN Results starting from classical features.*

Channel	Feature	Image	Accuracy	Precision			Recall		
				Nan	noM	yesM	Nan	noM	yesM
Magenta	Hm + Hf	Histeq	0.627	0.847	0.554	0.561	0.671	0.831	0.333
		CLAHE	0.690	0.929	0.607	0.611	0.761	0.790	0.465
		LoG	0.352	0.388	0.415	0.173	0.245	0.513	0.206
		DWT	0.358	0.392	0.410	0.216	0.283	0.464	0.270
	Hm + Hf (CLAHE) + Hi	Hi(LoG)	0.682	0.914	0.642	0.557	0.766	0.632	0.658
		Hi (DWT)	0.701	0.945	0.609	0.640	0.770	0.812	0.495
Green	Hm + Hf	Histeq	0.574	0.716	0.528	0.513	0.613	0.658	0.432
		CLAHE	0.683	0.898	0.609	0.605	0.752	0.750	0.532
		LoG	0.348	0.384	0.414	0.193	0.234	0.487	0.262
		DWT	0.363	0.409	0.412	0.227	0.283	0.464	0.298
	Hm + Hf (CLAHE) + Hi	Hi(LoG)	0.687	0.927	0.601	0.619	0.748	0.787	0.505
		Hi (DWT)	0.693	0.954	0.601	0.622	0.752	0.779	0.527
Magenta + Green	Hm + Hf (CLAHE)	Hi(LoG)	0.707	0.886	0.647	0.627	0.851	0.762	0.502
		Hi (DWT)	0.676	0.857	0.627	0.565	0.866	0.687	0.789

Table 4.11 *E-score — ANN Results starting from classical features.*

Channel	Feature	Image	Accuracy	Precision	Recall
Magenta	Hm + Hf	Histeq	0.615	0.029	0.500
		CLAHE	0.633	0.190	0.417
		LoG	0.606	0.015	0.250
		DWT	0.709	0.043	0.562
	Hm + Hf (CLAHE) + Hi	Hi (LoG)	0.756	0.018	0.250
		Hi (DWT)	0.736	0.047	0.562
Green	Hm + Hf	Histeq	0.655	0.250	0.375
		CLAHE	0.726	0.093	0.655
		LoG	0.598	0.021	0.500
		DWT	0.711	0.056	0.750
	Hm + Hf (CLAHE) + Hi	Hi (LoG)	0.772	0.013	0.167
		Hi (DWT)	0.758	0.051	0.562
Magenta + Green	Hm + Hf (CLAHE)	Hi (LoG)	0.778	0.108	0.621
		Hi (DWT)	0.799	0.124	0.655

Table 4.12 *S-score* — ANN Results starting from classical features.

Channel	Feature	Image	Accuracy	Precision			Recall		
				Nan	noM	yesM	Nan	noM	yesM
Magenta	Hm + Hf	Histeq	0.795	0.346	0.838	0.939	0.110	0.709	0.110
		CLAHE	0.832	0.333	0.917	0.972	0.049	0.880	0.049
		LoG	0.603	0.250	0.294	0.789	0.014	0.199	0.014
		DWT	0.795	0.556	0.929	0.980	0.061	0.749	0.061
	Hm + Hf (CLAHE) + Hi	Hi(LoG)	0.586	0.107	0.276	0.710	0.043	0.213	0.043
		Hi (DWT)	0.848	0.476	0.917	0.972	0.122	0.886	0.122
Green	Hm + Hf	Histeq	0.760	0.538	0.742	0.937	0.171	0.526	0.171
		CLAHE	0.821	0.611	0.923	0.972	0.134	0.817	0.134
		LoG	0.606	0.056	0.310	0.775	0.014	0.209	0.014
		DWT	0.828	0.409	0.918	0.967	0.110	0.829	0.110
	Hm + Hf (CLAHE) + Hi	Hi(LoG)	0.588	0.079	0.295	0.692	0.043	0.232	0.043
		Hi (DWT)	0.828	0.450	0.919	0.967	0.110	0.829	0.110
Magenta + Green	Hm + Hf (CLAHE)	Hi(LoG)	0.841	0.847	0.558	0.894	0.940	0.253	0.905
		Hi (DWT)	0.835	0.833	0.516	0.900	0.947	0.168	0.911

Table 4.13 *C-score* — ANN Results starting from classical features.

Channel	Feature	Image	Accuracy	Precision	Recall
Magenta	Hm + Hf	Histeq	0.668	0.127	0.750
		CLAHE	0.703	0.146	0.795
		LoG	0.845	0.023	0.074
		DWT	0.652	0.041	0.370
	Hm + Hf (CLAHE) + Hi	Hi (LoG)	0.832	0.189	0.523
		Hi (DWT)	0.697	0.150	0.841
Green	Hm + Hf	Histeq	0.690	0.126	0.682
		CLAHE	0.742	0.156	0.727
		LoG	0.634	0.043	0.407
		DWT	0.676	0.032	0.259
	Hm + Hf (CLAHE) + Hi	Hi (LoG)	0.730	0.164	0.700
		Hi (DWT)	0.732	0.161	0.795
Magenta + Green	Hm + Hf (CLAHE)	Hi (LoG)	0.768	0.174	0.620
		Hi (DWT)	0.772	0.188	0.680

Table 4.14 *M-score* — ANN Results starting from CNN feature extraction.

Model	Accuracy	Precision			Sensitivity			Specificity			F1 Score			AUC		
		nan	noM	yesM	nan	noM	yesM	Nan	noM	yesM	nan	noM	yesM	Nan	noM	yesM
ReNet18	0.729	0.872	0.827	0.686	0.846	0.827	0.493	0.952	0.716	0.901	0.859	0.741	0.574	0.968	0.848	0.837
ResNet50	0.742	0.883	0.776	0.679	0.866	0.776	0.584	0.963	0.749	0.881	0.874	0.732	0.628	0.969	0.844	0.843
GoogLeNet	0.682	0.694	0.694	0.581	0.841	0.694	0.520	0.938	0.713	0.880	0.845	0.665	0.549	0.950	0.815	0.816
SqueezeNet	0.729	0.676	0.653	0.900	0.772	0.561	0.851	0.965	0.718	0.871	0.721	0.604	0.875	0.964	0.846	0.842
MobileNet	0.750	0.890	0.727	0.657	0.846	0.770	0.624	0.959	0.796	0.855	0.867	0.748	0.640	0.968	0.861	0.848

Table 4.15 *E-score* — ANN Results starting from CNN feature extraction.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC
ResNet18	0.866	0.162	0.552	0.879	0.250	0.807
ResNet50	0.828	0.139	0.621	0.837	0.226	0.839
GoogLeNet	0.798	0.113	0.586	0.806	0.189	0.782
SqueezeNet	0.824	0.124	0.552	0.840	0.202	0.784
MobileNet	0.813	0.128	0.621	0.821	0.212	0.777

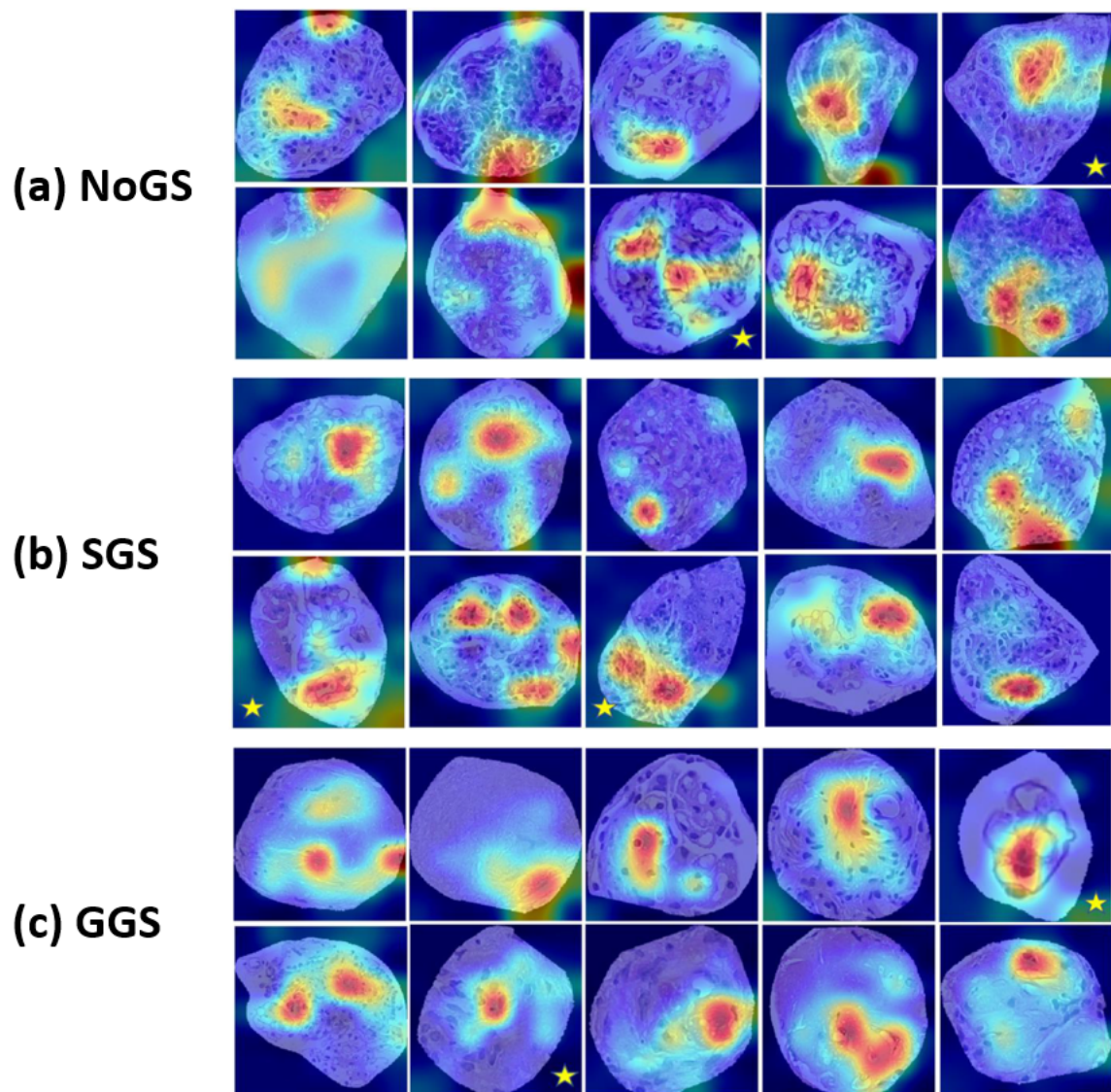


Fig. 4.26 *Grad-CAM* sample results for *S-score*. Examples are displayed in three blocks: (a) images from the NoGS class; (b) images from the SGS class; (c) images from the GGS class. Misclassified cases are marked with a star.

Table 4.16 S-score — ANN Results starting from CNN feature extraction.

Model	Accuracy	Precision			Sensitivity			Specificity			F1 Score			AUC		
		NoGS	SGS	GGs	NoGS	SGS	GGs	NoGS	SGS	GGs	NoGS	SGS	GGs	NoGS	SGS	GGs
ResNet18	0.838	0.833	0.475	0.948	0.958	0.200	0.875	0.669	0.966	0.985	0.891	0.282	0.910	0.905	0.805	0.986
ResNet50	0.866	0.890	0.573	0.929	0.927	0.453	0.935	0.791	0.953	0.982	0.908	0.506	0.932	0.928	0.858	0.989
ResNet50 + PCA	0.834	0.850	0.464	0.944	0.912	0.337	0.905	0.707	0.973	0.975	0.879	0.390	0.924	0.906	0.785	0.989
GoogLeNet	0.831	0.823	0.667	0.871	0.962	0.105	0.887	0.650	0.992	0.967	0.887	0.181	0.879	0.776	0.881	0.972
ResNet101	0.858	0.870	0.500	0.952	0.943	0.316	0.935	0.753	0.957	0.982	0.905	0.387	0.943	0.927	0.849	0.987
VGG16	0.848	0.851	0.567	0.884	0.949	0.179	0.952	0.730	0.974	0.967	0.897	0.272	0.917	0.924	0.841	0.990
AlexNet	0.841	0.830	0.524	0.916	0.969	0.116	0.905	0.677	0.962	0.978	0.894	0.190	0.910	0.890	0.803	0.979
InceptionV3	0.849	0.870	0.479	0.952	0.918	0.368	0.935	0.714	0.979	0.975	0.893	0.416	0.943	0.930	0.856	0.989
SqueezeNet	0.845	0.848	0.500	0.902	0.951	0.179	0.935	0.737	0.958	0.975	0.897	0.264	0.918	0.903	0.792	0.989
MobileNet	0.869	0.863	0.640	0.957	0.960	0.337	0.923	0.738	0.971	0.987	0.909	0.442	0.939	0.924	0.857	0.984
Retrain ResNet50	0.832	0.929	0.444	0.928	0.834	0.663	0.923	0.883	0.871	0.973	0.879	0.531	0.925	0.933	0.862	0.987

Table 4.17 *C-score* — ANN Results starting from CNN feature extraction.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC
ResNet18	0.806	0.230	0.760	0.809	0.354	0.864
ResNet50	0.782	0.206	0.740	0.785	0.322	0.839
GoogLeNet	0.694	0.149	0.720	0.692	0.247	0.817
SqueezeNet	0.790	0.231	0.860	0.772	0.364	0.891
MobileNet	0.758	0.185	0.720	0.761	0.294	0.837

4.3.4 Summary

The Oxford Classification for IgA glomerulonephritis is the most successful example of an evidence-based nephropathology classification system. The aim of the proposed machine learning approach is to develop a model for the automatic analysis of large biopsy cohorts with perfect reproducibility and minimal human effort.

A dataset of 3575 glomerular image crops from PAS sections of biopsies with IgA-GN was exploited. All crops were labelled by an expert nephropathologist for the M-score as *indeterminate*, *yes/no*, the E-scores as *yes/no*, the S-score as *global glomerulosclerosis/segmental glomerulosclerosis/no glomerulosclerosis*, the C-score as *yes/no*.

Several neural network architectures were trained for the M, E, S, C classifiers and compared the performance with standard metrics. The dataset was split into 64%/16%/20% for training, validation, and testing. Dataset disbalance was addressed with ADASYN. Different topologies of ANN were trained on top of both classical and deep CNN features extracted in a transfer learning paradigm for 500 epochs, exploiting the cross-entropy loss function.

The frequency of expert labels regarding M were 42%/30%/28% for *yes/nolindeterminate*, regarding E was 2%/98% for *yes/no*, regarding S was 23%/13%/64% for *global glomerulosclerosis/segmental glomerulosclerosis/no glomerulosclerosis*, regarding C was 6%/95% for *yes/no*. The best results were obtained with the transfer learning framework.

For the M lesion, the best AUC was 0.848 achieved by MobileNet. Corresponding F_1 -score, specificity, sensitivity, precision, and accuracy values were 0.640, 0.855, 0.624, 0.657, 0.750, respectively. The SqueezeNet model led to more balanced results, with an AUC of 0.848, an F_1 -score of 0.875, a specificity of 0.871, a sensitivity of 0.851, a precision of 0.900, and an accuracy of 0.729.

For the E lesion, the best AUC was 0.839 attained by ResNet50. Corresponding accuracy, precision, sensitivity, specificity, and F_1 -score values were 0.828, 0.139, 0.621, 0.837, 0.226,

and 0.839, respectively. ResNet18 managed to achieve slightly better accuracy, but with a slightly lower AUC. Indeed, the metric values were as follows: AUC of 0.807, F_1 -score of 0.226, specificity of 0.837, sensitivity of 0.621, precision of 0.139, and accuracy of 0.866.

For the S lesion, the best AUC for the detection of segmental glomerulosclerosis was demonstrated by GoogLeNet, with a value of 0.881. Instead, for the detection of global glomerulosclerosis, VGG-16 managed to achieve an AUC of 0.990. Retraining the ResNet50 model has consented to obtain balanced results. The overall accuracy is 0.832. Precision, sensitivity, specificity, F_1 -score, and AUC were 0.444, 0.663, 0.871, 0.531, and 0.862, respectively, for segmental glomerulosclerosis; 0.928, 0.923, 0.973, 0.925 and 0.987, respectively, for global glomerulosclerosis.

For the C lesion, the SqueezeNet displayed the overall best results. It managed to achieve an AUC of 0.891, an F_1 -score of 0.364, a specificity of 0.772, a sensitivity of 0.860, a precision of 0.231, and an accuracy of 0.790.

Hence, features extracted from modern deep network architectures, pre-trained on ImageNet, deliver promising results for the replication of expert labels for the glomerular components of the Oxford classification. However, larger training sets are needed to overcome the disbalance of rare lesions like E and C in randomly generated datasets.

Ongoing work includes the investigation of the role of Generative Adversarial Networks for generating such glomerular crops with rare lesions, exploiting the truncation parameter of StyleGAN2 ADA [449].

As part of future work, an automatic segmentation model will be deployed to generate larger datasets from more institutions. Finally, an end-to-end protocol, capable to provide M, E, S, C outputs for the input of unsegmented whole slide images, will be implemented and validated.

Chapter 5

Conclusion

The research works carried out for this thesis were devoted to the conceptualization, implementation and evaluation of novel systems that can assist physicians in the routine practice, offering an objective methodology to solve real medical problems.

In the proposed pipelines, different objectives were considered, including enhancement of the clinical performance, disposability and usability of the methodology, and feasibility of the integration into the daily clinical practice. Moreover, the investigation of principles behind the developed systems, by exploiting perceptive or mathematical explainability techniques, has been carried out when it has been possible.

Applications concerned a wide range of medical tasks, including data coming from Radiology and Digital Pathology setups. In the studies carried out for this thesis, data were provided by local institutions or come from public repositories. Physicians have contributed to the validation of most of the proposed systems.

A depth study of the state-of-the-art in the considered domains is presented in Chapter 2, with a focus on Deep Learning, Radiomics, and Pathomics, the emerging paradigms which are revolutionizing the medical practice in our days.

The proposed systems are detailed in Chapters 3 and 4 for Radiology and Digital Pathology, respectively. Each section details the materials and methods, the design decisions, as well as the results, and draws a summary concerning the study. Results have been compared with other works found in the state-of-the-art analysis. Pros and cons of the proposed approaches have been summarized and discussed even in a comparison with the existing others.

Chapter 3 describes four case studies in Radiology: liver segments classification for surgical planning, vertebrae segmentation and identification, prostate image segmentation and registration to achieve fusion biopsy, and lung COVID-19 lesions characterization.

The framework proposed in Section 3.1 exploits two novel formulations for the focal Dice loss, which permit to obtain a reliable classification of the liver segments, according to the Couinaud model, which has a fundamental role in modern hepatic surgery, since it allows to reduce the complications of more extensive parenchyma resections. The best model was capable to achieve an average Dice coefficient higher than 83% and an average accuracy higher than 98%.

In Section 3.2, a novel approach is proposed for vertebrae segmentation and identification. The implemented pipeline allows to achieve accurate results, being the average multi-class Dice coefficient higher than 90%, with a fast and straightforward approach. Moreover, the vertebrae identification stage successfully exploits unsupervised learning and therefore does not require training data but only a moderate user interaction. The clinical value of this method concerns the possibility to augment the functionalities of navigators exploited in minimally invasive procedures for spine surgery.

Section 3.3 describes a setup for fusion prostate biopsy, which involves segmentation from TRUS images with the technique of deformable superellipses, segmentation from MRI with the nnU-Net framework, and registration among the two imaging modalities. Fusion prostate biopsy provides a more reliable and accurate procedure than traditional prostate biopsy which extracts 10 to 12 tissue samples from predetermined prostate anatomical regions, without considering the region annotated in the MRI by the radiologist. Segmentation results are promising, being the Dice coefficient higher than 88% and 87% for the MRI and TRUS images, respectively. Image registration, which is mandatory to correctly perform image fusion, shows a Dice coefficient higher than 91% for all the considered cases.

Section 3.4 outlines a workflow devoted to characterizing lung parenchyma and lesions for COVID-19 patients. Indeed, one of the worst complications of COVID-19 is the onset of pulmonary thromboembolism. CT scans are the clinical routine procedure to assess the severity of lung lesions but do not have the same sensitivity as CTA scans to diagnose PTE. In this study, a segmentation pipeline, based on V-Net, is developed, and a radiomic signature is extracted to characterize the patients with a higher risk of PTE onset. The segmentation stage achieved a Dice coefficient higher than 97%, whereas 9 uncorrelated radiomic features demonstrate strong statistical significance, achieving a corrected p -value lesser than 0.01.

Chapter 4 inscribes three case studies in Digital Pathology: stain color normalization for CRC tissue classification, nuclei detection from histopathological images, and a Computer-Aided Diagnosis System for IgA glomerulonephritis Oxford Classification.

Section 4.1 describes the role of Unsupervised Image-to-Image Translation frameworks in stain color normalization. Several GAN models are compared between themselves and

with respect to traditional image processing color normalization techniques. Novel insights are drawn, and a novel metric, PPQ, based on the pathologist's perceptive quality assessment of images, is introduced. Particularly, PPQ shows correlation to FID and LPIPS, and FID to classification metrics, thus showing interesting relationships between the quality of generated images, both from a mathematical and an expert pathologist perspective, and the accuracy of downstream classifiers. The best performing normalization technique, FastCUT, paired with DenseNet201 feature extraction, manages to achieve an accuracy of more than 84% on the locally collected validation set.

Section 4.2 presents a novel method, NDG-CAM, for performing the detection of nuclei by exploiting Semantic Segmentation networks and Grad-CAM. Even without the need to use specialized loss functions or architectures, it consented to achieve satisfactory results in the detection task, comparable to or even better than more complex training setups. When the method is combined with the Mask R-CNN instance segmentation architecture, results exceed the state-of-the-art methods for nuclei detection. In detail, the NDG-CAM method achieves a Dice coefficient of 82.4% and 91.0% on the publicly available and the locally collected validation sets, respectively. The combined model obtains a very impressive Dice coefficient of 88.4% and 91.4% on the same validation datasets, showing the effectiveness of combining diverse techniques for the nuclei detection task.

Section 4.3 proposes a CAD for performing the Oxford Classification for IgA glomerulonephritis. The aim of the proposed DL-based approach is to develop a model for the automatic analysis of large biopsy cohorts with perfect reproducibility and minimal human effort. Features extracted from modern deep network architectures, pretrained on ImageNet, deliver promising results for the replication of expert labels for the glomerular components of the Oxford classification. However, larger training sets are needed to overcome the disbalance of rare lesions like E and C in randomly generated datasets. As future work, an automatic segmentation model will be deployed to generate larger datasets from more institutions. Finally, an end-to-end protocol, capable to provide M, E, S, C outputs for the input of unsegmented whole slide images, will be implemented and validated.

In conclusion, in this thesis, several technical contributions have been proposed. They include the design of innovative algorithms for tasks that are clinically relevant, the enhancement of existing workflow, statistical analysis and visualization of data coming from new cohorts, and the realization of CDSSs that fuse existing techniques in a novel and effective way.

The proposed approaches, in all the case studies, were compared to existing methodologies and validated with the support of expert clinicians to enhance the validity of the systems.

XAI techniques have been employed, when needed, to have a better understanding of the developed systems' behavior.

Radiomics and Pathomics are shifting the radiologists' and pathologists' traditional evaluation processes by enabling computational techniques to objectively assess tissues and lesions, providing a numerical characterization of the diseases' phenotypes.

In the future, computational methodologies will play an increasingly important role in the clinical practice, supporting the decision-making process, and will add the possibility to discover novel and effective biomarkers, especially in data-driven disciplines such as radiology and histopathology.

My Publications

1. **Nicola Altini**, Tommaso Maria Marvulli, Francesco Alfredo Zito, Mariapia Caputo, Stefania Tommasi, Amalia Azzariti, Antonio Brunetti, Berardino Prencipe, Eliseo Mattioli, Simona De Summa and Vitoantonio Bevilacqua. *Comparison between Image Processing and CycleGAN-based Stain Color Normalization for Colorectal Cancer Tissue Classification. Submitted to Computer Methods and Programs in Biomedicine 2022.*
2. Gian Maria Zaccaria, Maria Carmela Vegliante, Giuseppe Mezzolla, Marianna Stranieri, Giacomo Volpe, **Nicola Altini**, Grazia Gargano, Anita Susanna Pappagallo, Antonella Bucci, Flavia Esposito, Giuseppina Opinto, Felice Clemente, Antonio Negri, Paolo Mondelli, Maria Stella De Candia, Vitoantonio Bevilacqua, Attilio Guarini and Sabino Ciavarella. *A decision-tree approach to stratify DLBCL risk based on stromal and immune microenvironment determinants. Submitted to HemaSphere 2022.*
3. **Nicola Altini**, Antonio Brunetti, Emilia Puro, Maria Giovanna Taccogna, Concetta Saponaro, Francesco Alfredo Zito, Simona De Summa and Vitoantonio Bevilacqua. *NDG-CAM: Nuclei Detection in Histopathology Images with Semantic Segmentation Networks and Grad-CAM. Bioengineering 2022, 9, 475. doi:10.3390/bioengineering9090475.*
4. **Nicola Altini**, Antonio Brunetti, Valeria Pia Napoletano, Francesca Girardi, Emanuela Allegretti, Sardar Mehboob Hussain, Gioacchino Brunetti, Vito Triggiani, Vitoantonio Bevilacqua and Domenico Buongiorno. *A Fusion Biopsy Framework for Prostate Cancer Based on Deformable Superellipses and nnU-Net. Bioengineering 2022, 9, 343. doi:10.3390/bioengineering9080343.*
5. Sardar Mehboob Hussain, Domenico Buongiorno, **Nicola Altini**, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua and Antonio Brunetti. *Shape based Breast Lesion Classification using Digital Tomosynthesis Images:*

-
- the role of Explainable Artificial Intelligence*. Appl. Sci. 2022, 12, 6230. doi:10.3390/app12126230.
6. Antonio Brunetti, **Nicola Altini**, Domenico Buongiorno, Emilio Garolla, Fabio Corallo, Matteo Gravina, Vitoantonio Bevilacqua and Berardino Prencipe. *A Machine Learning and Radiomics Approach in Lung Cancer for Predicting Histological Subtype*. Appl. Sci. 2022, 12, 5829. doi:10.3390/app12125829.
 7. Berardino Prencipe, **Nicola Altini**, Giacomo Donato Cascarano, Antonio Brunetti, Andrea Guerriero and Vitoantonio Bevilacqua. *Focal Dice Loss-based V-Net for Liver Segments Classification*. Appl. Sci. 2022, 12, 3247. doi:10.3390/app12073247.
 8. **Nicola Altini**, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane and Vitoantonio Bevilacqua. *Liver, Kidney and Spleen Segmentation from CT scans and MRI with Deep Learning: A Survey*. Neurocomputing, Volume 490, 2022, Pages 30-53, ISSN 0925-2312, doi:10.1016/j.neucom.2021.08.157.
 9. **Nicola Altini**, Antonio Brunetti, Stefano Mazzoleni, Fabrizio Moncelli, Ilenia Zagaria, Berardino Prencipe, Erika Lorusso, Enrico Buonamico, Giovanna Elisiana Carpagnano, Davide Fiore Bavaro, Mariacristina Poliseo, Annalisa Saracino, Annalisa Schirinzi, Riccardo Laterza, Francesca Di Serio, Alessia D’Introno, Francesco Pesce and Vitoantonio Bevilacqua. *Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis based on Hematochemical Parameters*. Sensors 2021, 21, 8503. doi:10.3390/s21248503.
 10. Vitoantonio Bevilacqua, **Nicola Altini**, Berardino Prencipe, Antonio Brunetti, Laura Villani, Antonello Sacco, Chiara Morelli, Michele Ciaccia and Arnaldo Scardapane. *Lung Segmentation and Characterization in COVID-19 Patients for Assessing Pulmonary Thromboembolism: an Approach based on Deep Learning and Radiomics*. Electronics 2021, 10, 2475. doi:10.3390/electronics10202475.
 11. **Nicola Altini**, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibillano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, Iliana Sabina Tatò, Gioacchino Brunetti, Vito Triggiani and Vitoantonio Bevilacqua. *Segmentation and Identification of Vertebrae*

-
- in CT Scans Using CNN, k-Means Clustering and k-NN*. Informatics 2021, 8, 40. doi:10.3390/informatics8020040.
12. Fabio Tarricone, Antonio Brunetti, Domenico Buongiorno, **Nicola Altini**, Vitoantonio Bevilacqua, Antonio Del Vecchio and Flavia Petrillo. *Intelligent Neonatal Sepsis Early Diagnosis System for Very Low Birth Weight Infants*. Appl. Sci. 2021, 11, 404. doi:10.3390/app11010404.
 13. **Nicola Altini**, Giacomo Donato Cascarano, Antonio Brunetti, Irio De Feudis, Domenico Buongiorno, Michele Rossini, Francesco Pesce, Loreto Gesualdo and Vitoantonio Bevilacqua. *A Deep Learning Instance Segmentation Approach for Global Glomerulosclerosis Assessment in Donor Kidney Biopsies*. Electronics 2020, 9, 1768. doi:10.3390/electronics9111768.
 14. **Nicola Altini**, Giacomo Donato Cascarano, Antonio Brunetti, Francescomaria Marino, Maria Teresa Rocchetti, Silvia Matino, Umberto Venere, Michele Rossini, Francesco Pesce, Loreto Gesualdo and Vitoantonio Bevilacqua. *Semantic Segmentation Framework for Glomeruli Detection and Classification in Kidney Histological Sections*. Electronics 2020, 9, 503. doi:10.3390/electronics9030503.
 15. **Nicola Altini**, Tommaso Maria Marvulli, Mariapia Caputo, Eliseo Mattioli, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Stefania Tommasi, Vitoantonio Bevilacqua, Simona De Summa and Francesco Alfredo Zito. *Multi-class Tissue Classification in Colorectal Cancer with Handcrafted and Deep Features*. In: Huang DS., Jo KH., Li J., Gribova V., Bevilacqua V. (eds) Intelligent Computing Theories and Application. ICIC 2021. Lecture Notes in Computer Science, vol 12836. Springer, Cham. doi:10.1007/978-3-030-84522-3_42.
 16. Berardino Prencipe, **Nicola Altini**, Giacomo Donato Cascarano, Andrea Guerriero and Antonio Brunetti. *A Novel Approach based on Region Growing Algorithm for Liver and Spleen Segmentation from CT Scans*. In: Huang DS., Bevilacqua V., Hussain A. (eds) Intelligent Computing Theories and Application. ICIC 2020. Lecture Notes in Computer Science, vol 12463. Springer, Cham. doi:10.1007/978-3-030-60799-9_35.
 17. **Nicola Altini**, Berardino Prencipe, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane and Giacomo Donato Cascarano. *A Tversky Loss-based Convolutional Neural Network for Liver Vessels Segmentation*. In: Huang DS., Bevilacqua

- V., Hussain A. (eds) Intelligent Computing Theories and Application. ICIC 2020. Lecture Notes in Computer Science, vol 12463. Springer, Cham. doi:10.1007/978-3-030-60799-9_30.
18. Antonio Brunetti, Mariapia Caputo, Tommaso Maria Marvulli, Giacomo Donato Cascarano, **Nicola Altini**, Simona De Summa, Vitoantonio Bevilacqua. *Combining autoencoder and artificial neural network for classifying colorectal cancer stages*. Proceedings of VII Congress of the National Group of Bioengineering (GNB), June 9th-11th 2021, Trieste, Italy.

References

- [1] Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263, 2019.
- [2] Inke R Koenig, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V Kopp. What is precision medicine? *European respiratory journal*, 50(4), 2017.
- [3] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10, 2020.
- [4] Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200, 2018.
- [5] Dorin Comaniciu, Klaus Engel, Bogdan Georgescu, and Tommaso Mansi. Shaping the future through innovations: From medical imaging to precision medicine, 2016.
- [6] Roland Eils and Chaitanya Athale. Computational imaging in cell biology. *The Journal of cell biology*, 161(3):477–481, 2003.
- [7] E Sandeep Kumar and Pappu Satya Jayadev. Deep learning for clinical decision support systems: A review from the panorama of smart healthcare. *Deep Learning Techniques for Biomedical and Health Informatics*, pages 79–99, 2020.
- [8] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G P M Van Stiphout, Patrick Granton, Catharina M L Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.
- [9] Rajarsi Gupta, Tahsin Kurc, Ashish Sharma, Jonas S Almeida, and Joel Saltz. The emergence of pathomics. *Current Pathobiology Reports*, 7(3):73–84, 2019.
- [10] Arman Rahmim, C Ross Schmidtlein, Andrew Jackson, Sara Sheikhabaei, Charles Marcus, Saeed Ashrafinia, Madjid Soltani, and Rathan M Subramaniam. A novel metric for quantification of homogeneous and heterogeneous tumors in pet for enhanced clinical outcome prediction. *Physics in Medicine & Biology*, 61(1):227, 2015.

- [11] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J.W.L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J.R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkens, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, V. Lisanne van Dijk, Joost van Griethuyssen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. ISSN 15271315. doi: 10.1148/radiol.2020191145.
- [12] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Cristiana Fanciullo, Alessio Giuseppe Morganti, and Massimo Bellomi. Radiomics: the facts and the challenges of image analysis. *European radiology experimental*, 2(1):1–8, 2018.
- [13] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, 2012. ISSN 0730725X. doi: 10.1016/j.mri.2012.06.010.
- [14] Hugo J W L Aerts, Emmanuel Rios Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):1–9, 2014.
- [15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [16] Nicola Altini, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane, and Vitoantonio Bevilacqua. Liver, kidney and spleen segmentation from ct scans and mri with deep learning: A survey. *Neurocomputing*, 490:30–53, 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.08.157.
- [17] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 8 2018. ISSN 1869-4101. doi: 10.1007/s13244-018-0639-9.

- [18] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Francescomaria Marino, Maria Teresa Rocchetti, Silvia Matino, Umberto Venere, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. Semantic segmentation framework for glomeruli detection and classification in kidney histological sections. *Electronics*, 9(3):503, 3 2020. ISSN 2079-9292. doi: 10.3390/electronics9030503.
- [19] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Irio De Feudis, Domenico Buongiorno, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies. *Electronics*, 9(11):1768, 2020. ISSN 2079-9292. doi: 10.3390/electronics9111768.
- [20] Nicola Altini, Tommaso Maria Marvulli, Mariapia Caputo, Eliseo Mattioli, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Stefania Tommasi, Vitoantonio Bevilacqua, Simona De Summa, and Francesco Alfredo Zito. Multi-class tissue classification in colorectal cancer with handcrafted and deep features. In De-Shuang Huang, Kang-Hyun Jo, Jianqiang Li, Valeriya Gribova, and Vitoantonio Bevilacqua, editors, *Intelligent Computing Theories and Application*, pages 512–525, Cham, 2021. Springer International Publishing. ISBN 978-3-030-84522-3. doi: 10.1007/978-3-030-84522-3_42.
- [21] Thomas de Bel, John-Melle Bokhorst, Jeroen van der Laak, and Geert Litjens. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70:102004, 2021.
- [22] Nicola Altini, Antonio Brunetti, Emilia Puro, Maria Giovanna Taccogna, Concetta Saponaro, Francesco Alfredo Zito, Simona De Summa, and Vitoantonio Bevilacqua. Ndgcam: Nuclei detection in histopathology images with semantic segmentation networks and gradcam. *Bioengineering*, 9(9):475, 2022. ISSN 2306-5354. doi: 10.3390/bioengineering9090475.
- [23] Henning Höfener, André Homeyer, Nick Weiss, Jesper Molin, Claes F. Lundström, and Horst K. Hahn. Deep learning nuclei detection: A simple approach can deliver state-of-the-art results. *Computerized Medical Imaging and Graphics*, 70:43–52, 2018. ISSN 0895-6111. doi: <https://doi.org/10.1016/j.compmedimag.2018.08.010>. URL <https://www.sciencedirect.com/science/article/pii/S0895611118300806>.
- [24] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):3877, 2020. ISSN 20411723. doi: 10.1038/s41467-020-17678-4.
- [25] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting

- survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):1–22, 2019. ISSN 15491676. doi: 10.1371/journal.pmed.1002730.
- [26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances In Neural Information Processing Systems*, 2012. ISSN 10495258. doi: <http://dx.doi.org/10.1016/j.protcy.2014.09.007>.
- [27] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, 8:716007, 2021.
- [28] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.
- [29] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.101552.
- [30] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020.
- [31] Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua, and Antonio Brunetti. Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence. *Applied Sciences*, 12(12), 2022. ISSN 2076-3417. doi: 10.3390/app12126230.
- [32] Thomas Ploug and Søren Holm. The four dimensions of contestable ai diagnostics—a patient-centric approach to explainable ai. *Artificial Intelligence in Medicine*, 107: 101901, 2020.
- [33] Nicola Altini, Antonio Brunetti, Valeria Pia Napoletano, Francesca Girardi, Emanuela Allegretti, Sardar Mehboob Hussain, Gioacchino Brunetti, Vito Triggiani, Vitoantonio Bevilacqua, and Domenico Buongiorno. A fusion biopsy framework for prostate cancer based on deformable superellipses and nnu-net. *Bioengineering*, 9(8), 2022. ISSN 2306-5354. doi: 10.3390/bioengineering9080343.
- [34] Nicola Altini, Antonio Brunetti, Stefano Mazzoleni, Fabrizio Moncelli, Ilenia Zagaria, Berardino Prencipe, Erika Lorusso, Enrico Buonamico, Giovanna Elisiana Carpagnano, Davide Fiore Bavaro, Mariacristina Poliseno, Annalisa Saracino, Annalisa Schirinzi, Riccardo Laterza, Francesca Di Serio, Alessia D’Introno, Francesco Pesce, and Vitoantonio Bevilacqua. Predictive machine learning models and survival analysis for covid-19 prognosis based on hematochemical parameters. *Sensors*, 21(24), 2021. ISSN 1424-8220. doi: 10.3390/s21248503.

- [35] Berardino Prencipe, Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Andrea Guerriero, and Vitoantonio Bevilacqua. Focal dice loss-based v-net for liver segments classification. *Applied Sciences*, 12(7), 2022. ISSN 2076-3417. doi: 10.3390/app12073247.
- [36] Nicola Altini, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibilano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, Iliara Sabina Tatò, Gioacchino Brunetti, Vito Triggiani, and Vitoantonio Bevilacqua. Segmentation and identification of vertebrae in ct scans using cnn, k-means clustering and k-nn. *Informatics*, 8(2), 2021. ISSN 2227-9709. doi: 10.3390/informatics8020040.
- [37] Vitoantonio Bevilacqua, Nicola Altini, Berardino Prencipe, Antonio Brunetti, Laura Villani, Antonello Sacco, Chiara Morelli, Michele Ciaccia, and Arnaldo Scardapane. Lung segmentation and characterization in covid-19 patients for assessing pulmonary thromboembolism: An approach based on deep learning and radiomics. *Electronics*, 10(20):2475, 2021. ISSN 2079-9292. doi: 10.3390/electronics10202475.
- [38] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 12 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541.
- [39] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 4 2017. ISSN 09252312. doi: 10.1016/j.neucom.2016.12.038.
- [40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning: Machine learning book. page 785, 2016.
- [41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14, 2014.
- [42] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1):818–833, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10590-1_53.
- [43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. volume 07-12-June-2015, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298594.
- [44] Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, Liru Guo, and Tao Liu. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1303–1314, 2018. ISSN 10518215. doi: 10.1109/TCSVT.2017.2654543.

- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. pages 1026–1034. IEEE, 12 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123.
- [46] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. (i): 1–10, 2018.
- [47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 12 2015.
- [48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 12 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90.
- [49] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(1995):60–88, 12 2017. ISSN 13618415. doi: 10.1016/j.media.2017.07.005.
- [50] Shujian Deng, Xin Zhang, Wen Yan, Eric I Chang, Yubo Fan, Maode Lai, Yan Xu, et al. Deep learning in digital pathology image analysis: a survey. *Frontiers of medicine*, 14 (4):470–487, 2020.
- [51] Massimo Salvi, U Rajendra Acharya, Filippo Molinari, and Kristen M Meiburger. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128:104129, 2021.
- [52] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. ISSN 20010370. doi: 10.1016/j.csbj.2018.01.001.
- [53] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. ISSN 01628828. doi: 10.1109/TPAMI.2015.2389824.
- [55] Teresa Araujo, Guilherme Aresta, Eduardo Castro, Jose Rouco, Paulo Aguiar, Catarina Eloy, Antonio Polonia, and Aurelio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE*, 12(6):1–14, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0177544.

- [56] Nicola Altini, Berardino Prencipe, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane, and Giacomo Donato Cascarano. A tversky loss-based convolutional neural network for liver vessels segmentation. In De-Shuang Huang, Vitoantonio Bevilacqua, and Abir Hussain, editors, *Intelligent Computing Theories and Application*, pages 342–354, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60799-9. doi: 10.1007/978-3-030-60799-9_30.
- [57] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- [58] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11141 LNCS:270–279, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01424-7_27.
- [59] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8(1):1–11, 2018. ISSN 20452322. doi: 10.1038/s41598-018-21758-3.
- [60] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. ISSN 1546170X. doi: 10.1038/s41591-019-0508-1.
- [61] J Fu, K Singhrao, M Cao, V Yu, A P Santhanam, Y Yang, M Guo, A C Raldow, D Ruan, and J H Lewis. Generation of abdominal synthetic cts from 0.35t mr images using generative adversarial networks for mr-only liver radiotherapy. *Biomedical Physics and Engineering Express*, 6(1), 2020. doi: 10.1088/2057-1976/ab6e1f. cited By 1.
- [62] Alona Levy-jurgenson. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. pages 1–16, 2020.
- [63] Silvia Cascianelli, Raquel Bello-Cerezo, Francesco Bianconi, Mario L Fravolini, Mehdi Belal, Barbara Palumbo, and Jakob N Kather. Dimensionality reduction strategies for cnn-based classification of histopathological images. In *Intelligent interactive multimedia systems and services 2017*, volume 1, pages 21–30. 2018. ISBN 331959480X. doi: 10.1007/978-3-319-59480-4_3.
- [64] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

- [65] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. pages 2261–2269, 7 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.243.
- [66] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [67] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019. ISSN 18728286. doi: 10.1016/j.neucom.2019.02.003.
- [68] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017. ISSN 17518520. doi: 10.1007/s11633-017-1053-3.
- [69] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing Journal*, 70:41–65, 2018. ISSN 15684946. doi: 10.1016/j.asoc.2018.05.018.
- [70] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2644615.
- [71] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. pages 565–571. IEEE, 10 2016. ISBN 978-1-5090-5407-7. doi: 10.1109/3DV.2016.79.
- [72] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28.
- [73] Ozgun Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:424–432, 6 2016. ISSN 16113349. doi: 10.1007/978-3-319-46723-8_49.
- [74] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G. Elmore, and Linda Shapiro. Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11071 LNCS:893–901, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00934-2_99.

- [75] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. 11 2014.
- [76] Liang-Chieh Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 6 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2699184.
- [77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 37, pages 346–361. 2014. doi: 10.1007/978-3-319-10578-9_23.
- [78] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. 12 2014.
- [79] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.
- [80] Liang-Chieh Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:833–851, 2 2018. ISSN 16113349. doi: 10.1007/978-3-030-01234-2_49.
- [81] Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. A 2d dilated residual u-net for multi-organ segmentation in thoracic ct. *CEUR Workshop Proceedings*, 2349:2–5, 2019. ISSN 16130073.
- [82] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Leyi Wei, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. (October): 1–13, 11 2018.
- [83] Pulkit Kumar, Pravin Nagar, Chetan Arora, and Anubha Gupta. U-segnet: Fully convolutional neural network based automated brain tissue segmentation tool. *Proceedings - International Conference on Image Processing, ICIP*, (August):3503–3507, 2018. ISSN 15224880. doi: 10.1109/ICIP.2018.8451295.
- [84] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *Informatik aktuell*, page 22, 2019. ISSN 1431472X. doi: 10.1007/978-3-658-25326-4_7.

- [85] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. *Unet++: A nested u-net architecture for medical image segmentation*, volume 11045 LNCS. Springer International Publishing, 2018. ISBN 9783030008888. doi: 10.1007/978-3-030-00889-5_1. cited By 180.
- [86] Chen Shen, Fausto Milletari, Holger R. Roth, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. Improving v-nets for multi-class abdominal organ segmentation. page 10. SPIE, 3 2019. ISBN 9781510625457. doi: 10.1117/12.2512790.
- [87] X Li, H Chen, X Qi, Q Dou, C.-W. Fu, and P.-A. Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018. doi: 10.1109/TMI.2018.2845918. cited By 221.
- [88] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. ISSN 1558254X. doi: 10.1109/TMI.2019.2959609. cited By 12.
- [89] Liangliang Liu, Jianhong Cheng, Quan Quan, Fang Xiang Wu, Yu Ping Wang, and Jianxin Wang. A survey on u-shaped networks in medical image segmentations. *Neuro-computing*, 409:244–258, 2020. ISSN 18728286. doi: 10.1016/j.neucom.2020.05.070.
- [90] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [91] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [92] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [93] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [94] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [95] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. 2016.

- [96] Fabio Henrique Kiyoyiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. 2019:1–16, 2019.
- [97] Changhee Han, Kohei Murao, Shin’ichi Satoh, and Hideki Nakayama. Learning more with less: Gan-based medical image augmentation. pages 1–6, 2019. ISSN 0288-450X. doi: 10.11409/mit.37.137.
- [98] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5967–5976, 2017. doi: 10.1109/CVPR.2017.632.
- [99] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. pages 2223–2232, 2017.
- [100] He Huang, Philip S. Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets. *arXiv*, pages 1–17, 2018.
- [101] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018. ISSN 10535888. doi: 10.1109/MSP.2017.2765202.
- [102] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:6350–6367, 2019.
- [103] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I.Chao Chang, and Yan Xu. Mri cross-modality image-to-image translation. *Scientific Reports*, 10(1):1–18, 2020. ISSN 20452322. doi: 10.1038/s41598-020-60520-6.
- [104] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:417–425, 2017. ISSN 16113349. doi: 10.1007/978-3-319-66179-7_48.
- [105] Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, and Ivana Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE Transactions on Medical Imaging*, 36(12):2536–2545, 2017. ISSN 1558254X. doi: 10.1109/TMI.2017.2708987.
- [106] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9907 LNCS:649–666, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46487-9_40.
- [107] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. 11 2015.

- [108] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. 7 2012.
- [109] Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. Survey of dropout methods for deep neural networks. 4 2019.
- [110] Eunhee Kang, Hyun Jung Koo, Dong Hyun Yang, Joon Bum Seo, and Jong Chul Ye. Cycle-consistent adversarial denoising network for multiphase coronary ct angiography. *Medical Physics*, 46(2):550–562, 2019. ISSN 00942405. doi: 10.1002/mp.13284.
- [111] L. Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.101539. URL <https://doi.org/10.1016/j.media.2019.101539>.
- [112] Fernando Navarro, Christopher Watanabe, Suprosanna Shit, Anjany Sekuboyina, Jan C. Peeken, Stephanie E. Combs, and Bjoern H. Menze. Evaluating the Robustness of Self-Supervised Learning in Medical Imaging. pages 1–10, 2021. URL <http://arxiv.org/abs/2105.06986>.
- [113] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [114] Abhinav Dhere and Jayanthi Sivaswamy. Self-Supervised Learning for Segmentation. pages 1–4, 2021. URL <http://arxiv.org/abs/2101.05456>.
- [115] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal Self-supervised Learning for Medical Image Analysis. XX(Xx):661–673, 2021. doi: 10.1007/978-3-030-78191-0_51.
- [116] Minyoung Chung, Jingyu Lee, Minkyung Lee, Jeongjin Lee, and Yeong Gil Shin. Deeply self-supervised contour embedded neural network applied to liver segmentation. *Computer Methods and Programs in Biomedicine*, 192:105447, 2020. ISSN 18727565. doi: 10.1016/j.cmpb.2020.105447. URL <https://doi.org/10.1016/j.cmpb.2020.105447>.
- [117] Karin Stacke, Jonas Unger, Claes Lundström, and Gabriel Eilertsen. Learning representations with contrastive self-supervised learning for histopathology applications. *arXiv preprint arXiv:2112.05760*, 2021.
- [118] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [119] John D Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML ’01, page

- 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.
- [120] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pages 1–9, 2011.
- [121] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.
- [122] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. pages 2881–2890, 2017.
- [123] Marvin Teichmann and Roberto Cipolla. Convolutional crfs for semantic segmentation. *30th British Machine Vision Conference 2019, BMVC 2019*, 2020.
- [124] Miguel Monteiro, Mário A. T. Figueiredo, and Arlindo L. Oliveira. Conditional random fields as recurrent neural networks for 3d medical imaging segmentation. pages 1–11, 2018.
- [125] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H.S. Torr. Conditional random fields as recurrent neural networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:1529–1537, 2015. ISSN 15505499. doi: 10.1109/ICCV.2015.179.
- [126] Tobias Heimann, Brain van Ginneken, M.A. Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M.M. Cashman, Ying Chi, Andrés Córdova, B.M. Benoit M. Dawant, Márta Fidrich, J.D. Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar Kainmüller, R.I. Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, H.-P. Hans Peter Meinzer, Gábor Németh, D.S. Daniela S. Raicu, A.-M. Anne Mareike Rau, Eva M. E.M. van Rikxoort, Mikael Rousson, László Ruskó, K.A. Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. J.M. Waite, Andreas Wimmer, Ivo Wolf, Ying Chi, A. Cordova, B.M. Benoit M. Dawant, Márta Fidrich, J.D. Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, D. Kainmuller, R.I. Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, H.-P. Hans Peter Meinzer, G. Németh, D.S. Daniela S. Raicu, A.-M. Anne Mareike Rau, Eva M. E.M. van Rikxoort, Mikael Rousson, L. Rusko, K.A. Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. J.M. Waite, Andreas Wimmer, and Ivo Wolf. Comparison and evaluation of methods for liver segmentation from ct datasets.

- IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 8 2009. ISSN 0278-0062. doi: 10.1109/TMI.2009.2013851.
- [127] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, Samuel Kadoury, Tomasz Konopczynski, Miao Le, Chunming Li, Xiaomeng Li, Jana Lipková, John Lowengrub, Hans Meine, Jan Hendrik Moltz, Chris Pal, Marie Piraud, Xiaojuan Qi, Jin Qi, Markus Rempfler, Karsten Roth, Andrea Schenk, Anjany Sekuboyina, Eugene Vorontsov, Ping Zhou, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Felix Gruen, Georgios Kaissis, Fabian Lohöfer, Rickmer Braren, Julian Holch, Felix Hofmann, Wieland Sommer, Volker Heinemann, Colin Jacobs, Gabriel Efrain Humpire Mamani, Bram van Ginneken, Gabriel Chartrand, An Tang, Michal Drozdal, Avi Ben-Cohen, Eyal Klang, Marianne M. Amitai, Eli Konen, Hayit Greenspan, Johan Moreau, Alexandre Hostettler, Luc Soler, Refael Vivanti, Adi Szeskin, Naama Lev-Cohain, Jacob Sosna, Leo Joskowicz, and Bjoern H. Menze. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, pages 1–43, 1 2019.
- [128] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012.
- [129] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. *Proceedings - International Conference on Pattern Recognition*, (August):2366–2369, 2010. ISSN 10514651. doi: 10.1109/ICPR.2010.579.
- [130] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [131] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.
- [132] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [133] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [134] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [135] Lennox Hoyte, Wen Ye, Linda Brubaker, Julia R. Fielding, Mark E. Lockhart, Marta E. Heilbrun, Morton B. Brown, and Simon K. Warfield. Segmentations of mri images of

- the female pelvic floor: A study of inter- and intra-reader reliability. *Journal of Magnetic Resonance Imaging*, 33(3):684–691, 3 2011. ISSN 10531807. doi: 10.1002/jmri.22478.
- [136] Vitoantonio Bevilacqua, Giovanni Dimauro, Francescomaria Marino, Antonio Brunetti, Fabio Cassano, Antonio Di Maio, Enrico Nasca, Gianpaolo Francesco Trotta, Francesco Girardi, Angelo Ostuni, and Attilio Guarini. A novel approach to evaluate blood parameters using computer vision techniques. pages 1–6. IEEE, 5 2016. ISBN 978-1-4673-9172-6. doi: 10.1109/MeMeA.2016.7533760.
- [137] Vitoantonio Bevilacqua, Dario D’Ambruoso, Giovanni Mandolino, and Marco Suma. A new tool to support diagnosis of neurological disorders by means of facial expressions. pages 544–549. IEEE, 5 2011. ISBN 978-1-4244-9336-4. doi: 10.1109/MeMeA.2011.5966766.
- [138] Xiao-Feng Wang, De-Shuang Huang, and Huan Xu. An efficient local chan–vese model for image segmentation. *Pattern Recognition*, 43(3):603–618, 2010.
- [139] Zhong-Qiu Zhao, De-Shuang Huang, and Bing-Yu Sun. Human face recognition based on multi-features using neural networks committee. *Pattern Recognition Letters*, 25(12): 1351–1358, 2004. ISSN 0167-8655. doi: <https://doi.org/10.1016/j.patrec.2004.05.008>.
- [140] Ji-Xiang Du, De-Shuang Huang, Xiao-Feng Wang, and Xiao Gu. Shape recognition based on neural networks trained by differential evolution algorithm. *Neurocomputing*, 70(4-6):896–903, 2007.
- [141] Wen-Sheng Chen, Pong Chi Yuen, Jian Huang, and Dao-Qing Dai. Kernel machine-based one-parameter regularized fisher discriminant method for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(4):659–669, 2005.
- [142] Bo Li, Chun-Hou Zheng, and De-Shuang Huang. Locally linear discriminant embedding: An efficient method for face recognition. *Pattern Recognition*, 41(12):3813–3821, 2008.
- [143] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2021.102062>. URL <https://www.sciencedirect.com/science/article/pii/S1361841521001080>.
- [144] Zheyu Zhu, Yuzhe Lu, Ruining Deng, Haichun Yang, Agnes B. Fogo, and Yuankai Huo. Easierpath: An open-source tool for human-in-the-loop deep learning of renal pathology. In Jaime Cardoso, Hien Van Nguyen, Nicholas Heller, Pedro Henriques Abreu, Ivana Isgum, Wilson Silva, Ricardo Cruz, Jose Pereira Amorim, Vishal Patel, Badri Roysam, Kevin Zhou, Steve Jiang, Ngan Le, Khoa Luu, Raphael Sznitman, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, and Samaneh Abbasi, editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 214–222, Cham, 2020. Springer International Publishing.

- [145] Antonio Brunetti, Leonarda Carnimeo, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer-assisted frameworks for classification of liver, breast and blood neoplasias via neural networks: A survey based on medical images. *Neurocomputing*, 335:274–298, 3 2019. ISSN 09252312. doi: 10.1016/j.neucom.2018.06.080.
- [146] Vitoantonio Bevilacqua, Antonio Brunetti, Gianpaolo Francesco Trotta, Domenico De Marco, Marco Giuseppe Quercia, Domenico Buongiorno, Alessia D’Introno, Francesco Girardi, and Attilio Guarini. A novel deep learning approach in haematology for classification of leucocytes. *Smart Innovation, Systems and Technologies*, 103: 265–274, 2019. doi: 10.1007/978-3-319-95095-2_25.
- [147] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D’Anastasi, Wieland H. Sommer, Seyed Ahmad Ahmadi, and Björn H. Menze. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:415–423, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46723-8_48.
- [148] L M Prevedello, B S Erdal, J L Ryu, K J Little, M Demirer, S Qian, and R D White. Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology*, 285(3):923–931, 2017. doi: 10.1148/radiol.2017162664. cited By 77.
- [149] S Chen, J Qin, X Ji, B Lei, T Wang, D Ni, and J.-Z. Cheng. Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in ct images. *IEEE Transactions on Medical Imaging*, 36(3):802–814, 2017. doi: 10.1109/TMI.2016.2629462. cited By 53.
- [150] A Masood, B Sheng, P Li, X Hou, X Wei, J Qin, and D Feng. Computer-assisted decision support system in pulmonary cancer detection and stage classification on ct images. *Journal of Biomedical Informatics*, 79:117–128, 2018. doi: 10.1016/j.jbi.2018.01.005. cited By 52.
- [151] J Zhao, M Zhang, Z Zhou, J Chu, and F Cao. Automatic detection and classification of leukocytes using convolutional neural networks. *Medical and Biological Engineering and Computing*, 55(8):1287–1301, 2017. doi: 10.1007/s11517-016-1590-x. cited By 50.
- [152] V I Iglovikov, A Rakhlin, A A Kalinin, and A A Shvets. Paediatric bone age assessment using deep convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:300–308, 2018. doi: 10.1007/978-3-030-00889-5_34. cited By 32.
- [153] Y Zhao, H Li, S Wan, A Sekuboyina, X Hu, G Tetteh, M Piraud, and B Menze. Knowledge-aided convolutional neural network for small organ segmentation. *IEEE*

- Journal of Biomedical and Health Informatics*, 23(4):1363–1373, 2019. doi: 10.1109/JBHI.2019.2891526. cited By 30.
- [154] N Gessert, M Lutz, M Heyder, S Latus, D M Leistner, Y S Abdelwahed, and A Schlaefer. Automatic plaque detection in ivoct pullbacks using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 38(2):426–434, 2019. doi: 10.1109/TMI.2018.2865659. cited By 24.
- [155] S R Hashemi, S S M Salehi, D Erdogmus, S P Prabhu, S K Warfield, and A Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2019. doi: 10.1109/ACCESS.2018.2886371. cited By 21.
- [156] E Ahn, A Kumar, J Kim, C Li, D Feng, and M Fulham. X-ray image classification using domain transferred convolutional neural networks and local sparse spatial pyramid. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, volume 2016-June, pages 855–858, 2016. doi: 10.1109/ISBI.2016.7493400.
- [157] Vitoantonio Bevilacqua, Antonio Brunetti, Giacomo Donato Cascarano, Flavio Palmieri, Andrea Guerriero, and Marco Moschetta. A deep learning approach for the automatic detection and segmentation in autosomal dominant polycystic kidney disease based on magnetic resonance images. pages 643–649. 2018. doi: 10.1007/978-3-319-95933-7_73.
- [158] Vitoantonio Bevilacqua, Antonio Brunetti, Giacomo Donato Cascarano, Andrea Guerriero, Francesco Pesce, Marco Moschetta, and Loreto Gesualdo. A comparison between two semantic deep learning frameworks for the autosomal dominant polycystic kidney disease segmentation based on magnetic resonance images. *BMC Medical Informatics and Decision Making*, 19(S9):244, 12 2019. ISSN 1472-6947. doi: 10.1186/s12911-019-0988-4.
- [159] E Gibson, F Giganti, Y Hu, E Bonmati, S Bandula, K Gurusamy, B Davidson, S P Pereira, M J Clarkson, and D C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Transactions on Medical Imaging*, 37(8):1822–1834, 2018. doi: 10.1109/TMI.2018.2806309. cited By 108.
- [160] M H Hesamian, W Jia, X He, and P Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32(4):582–596, 2019. doi: 10.1007/s10278-019-00227-x. cited By 84.
- [161] H R Roth, L Lu, N Lay, A P Harrison, A Farag, A Sohn, and R M Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical Image Analysis*, 45:94–107, 2018. doi: 10.1016/j.media.2018.01.006. cited By 77.

- [162] H R Roth, A Farag, L Lu, E B Turkbey, and R M Summers. Deep convolutional networks for pancreas segmentation in ct imaging. volume 9413, 2015. doi: 10.1117/12.2081420. cited By 82.
- [163] X Dong, Y Lei, T Wang, M Thomas, L Tang, W J Curran, T Liu, and X Yang. Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Medical Physics*, 46(5):2157–2168, 2019. doi: 10.1002/mp.13458. cited By 58.
- [164] H R Roth, H Oda, X Zhou, N Shimizu, Y Yang, Y Hayashi, M Oda, M Fujiwara, K Misawa, and K Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018. doi: 10.1016/j.compmedimag.2018.03.001. cited By 55.
- [165] X Zhou, R Takayama, S Wang, T Hara, and H Fujita. Deep learning of the sectional appearances of 3d ct images for anatomical structure segmentation based on an fcnn voting method. *Medical Physics*, 44(10):5221–5233, 2017. doi: 10.1002/mp.12480. cited By 53.
- [166] N Tong, S Gou, S Yang, D Ruan, and K Sheng. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical Physics*, 45(10):4558–4567, 2018. doi: 10.1002/mp.13147. cited By 40.
- [167] X Dong, Y Lei, S Tian, T Wang, P Patel, W J Curran, A B Jani, T Liu, and X Yang. Synthetic mri-aided multi-organ segmentation on male pelvic ct using cycle consistent deep attention network. *Radiotherapy and Oncology*, 141:192–199, 2019. doi: 10.1016/j.radonc.2019.09.028. cited By 39.
- [168] Y Lei, T Wang, S Tian, X Dong, A B Jani, D Schuster, W J Curran, P Patel, T Liu, and X Yang. Male pelvic multi-organ segmentation aided by cbct-based synthetic mri. *Physics in Medicine and Biology*, 65(3), 2020. doi: 10.1088/1361-6560/ab63bb. cited By 37.
- [169] Bob D. de Vos, Jelmer M. Wolterink, Pim A. de Jong, Tim Leiner, Max A. Viergever, and Ivana Isgum. Convnet-based localization of anatomical structures in 3-d medical images. *IEEE Transactions on Medical Imaging*, 36(7):1470–1481, 7 2017. ISSN 0278-0062. doi: 10.1109/TMI.2017.2673121.
- [170] Xuesong Lu, Qinlan Xie, Yunfei Zha, and Defeng Wang. Fully automatic liver segmentation combining multi-dimensional graph cut with shape information in 3d ct images. *Scientific Reports*, 8(1):10700, 12 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-28787-y.
- [171] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8): 500–510, 8 2018. ISSN 1474-175X. doi: 10.1038/s41568-018-0016-5.

- [172] Morgan P. McBee, Omer A. Awan, Andrew T. Colucci, Comeron W. Ghobadi, Nadjia Kadom, Akash P. Kansagra, Srinidhi Tridandapani, and William F. Auffermann. Deep learning in radiology. *Academic Radiology*, 25(11):1472–1480, 2018. ISSN 18784046. doi: 10.1016/j.acra.2018.02.018.
- [173] Vitoantonio Bevilacqua, Antonio Brunetti, G.F. Gianpaolo Francesco Trotta, Giovanni Dimauro, Katarina Elez, Vito Alberotanza, and Arnaldo Scardapane. A novel approach for hepatocellular carcinoma detection and classification based on triphasic ct protocol. pages 1856–1863. IEEE, 6 2017. ISBN 978-1-5090-4601-0. doi: 10.1109/CEC.2017.7969527.
- [174] Antonio Brunetti, Giacomo Donato Casciaro, Iorio De Feudis, Marco Moschetta, Loreto Gesualdo, and Vitoantonio Bevilacqua. Detection and segmentation of kidneys from magnetic resonance images in patients with autosomal dominant polycystic kidney disease. pages 639–650. 2019. doi: 10.1007/978-3-030-26969-2_60.
- [175] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 5 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2535302.
- [176] J.-G. Lee, S Jun, Y.-W. Cho, H Lee, G B Kim, J B Seo, and N Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584, 2017. doi: 10.3348/kjr.2017.18.4.570. cited By 255.
- [177] G Chartrand, P M Cheng, E Vorontsov, M Drozdal, S Turcotte, C J Pal, S Kadoury, and A Tang. Deep learning: A primer for radiologists. *Radiographics*, 37(7):2113–2131, 2017. doi: 10.1148/rg.2017170077. cited By 211.
- [178] A Kumar, J Kim, D Lyndon, M Fulham, and D Feng. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 21(1):31–40, 2017. doi: 10.1109/JBHI.2016.2635663. cited By 156.
- [179] P Rajpurkar, J Irvin, R L Ball, K Zhu, B Yang, H Mehta, T Duan, D Ding, A Bagul, C P Langlotz, B N Patel, K W Yeom, K Shpanskaya, F G Blankenberg, J Seekins, T J Amrhein, D A Mong, S S Halabi, E J Zucker, A Y Ng, and M P Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the cheXnext algorithm to practicing radiologists. *PLoS Medicine*, 15(11), 2018. doi: 10.1371/journal.pmed.1002686. cited By 137.
- [180] S Chilamkurthy, R Ghosh, S Tanamala, M Biviji, N G Campeau, V K Venugopal, V Mahajan, P Rao, and P Warier. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet*, 392(10162):2388–2396, 2018. doi: 10.1016/S0140-6736(18)31645-3. cited By 126.

- [181] J H Thrall, X Li, Q Li, C Cruz, S Do, K Dreyer, and J Brink. Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, 15(3):504–508, 2018. doi: 10.1016/j.jacr.2017.12.026. cited By 100.
- [182] D D Miller and E W Brown. Artificial intelligence in medical practice: The question to the answer? *American Journal of Medicine*, 131(2):129–133, 2018. doi: 10.1016/j.amjmed.2017.10.035. cited By 100.
- [183] H.-C. Shin, K Roberts, L Lu, D Demner-Fushman, J Yao, and R M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. volume 2016-Decem, pages 2497–2506, 2016. doi: 10.1109/CVPR.2016.274. cited By 100.
- [184] A Tang, R Tam, A Cadrin-Chênevert, W Guest, J Chong, J Barfett, L Chepelev, R Cairns, J R Mitchell, M D Cicero, M G Poudrette, J L Jaremko, C Reinhold, B Gallix, B Gray, R Geis, T O’Connell, P Babyn, D Koff, D Ferguson, S Derkatch, A Bilbily, and W Shabana. Canadian association of radiologists white paper on artificial intelligence in radiology. *Canadian Association of Radiologists Journal*, 69(2):120–135, 2018. doi: 10.1016/j.carj.2018.02.002. cited By 99.
- [185] M L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018. doi: 10.1016/j.jacr.2017.12.028. cited By 98.
- [186] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- [187] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Scalable multi-label annotation. 2014.
- [188] Olga Russakovsky, Jia Deng, Zhiheng Huang, Alexander C Berg, and Li Fei-Fei. Detecting avocados to zucchinis: what have we done, and where are we going? 2013.
- [189] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. Crete, Greece, 2010.
- [190] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.
- [191] J Deng, K Li, M Do, H Su, and L Fei-Fei. Construction and analysis of a large scale image ontology. 2009.
- [192] Zhong-Qiu Zhao, Hervé Glotin, Zhao Xie, Jun Gao, and Xindong Wu. Cooperative sparse representation in two opposite directions for semi-supervised image annotation. *IEEE Transactions on Image Processing*, 21(9):4218–4231, 2012.

- [193] Ronald M. Summers. Progress in fully automated abdominal ct interpretation. *American Journal of Roentgenology*, 207(1):67–79, 7 2016. ISSN 0361-803X. doi: 10.2214/AJR.15.15996.
- [194] John Launchbury. A darpa perspective on artificial intelligence. *DARPA slides*, 2017.
- [195] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.
- [196] Andrea Barucci and Emanuele Neri. Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 10–12, 2020. ISSN 16197089. doi: 10.1007/s00259-020-04879-8.
- [197] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 493–501. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-030-00928-1_56.
- [198] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. pages 1765–1773, 2017.
- [199] M F Bobo, S Bao, Y Huo, Y Yao, J Virostko, A J Plassard, I Lyu, A Assad, R G Abramson, M A Hilmes, and B A Landman. Fully convolutional neural networks improve abdominal organ segmentation. volume 10574, 2018. doi: 10.1117/12.2293751. cited By 11.
- [200] V Sandfort, K Yan, P J Pickhardt, and R M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-52737-x. cited By 19.
- [201] S Chen, X Zhong, S Hu, S Dorn, M Kachelrieß, M Lell, and A Maier. Automatic multi-organ segmentation in dual-energy ct (dect) with dedicated 3d fully convolutional dect networks. *Medical Physics*, 47(2):552–562, 2020. doi: 10.1002/mp.13950. cited By 3.
- [202] Y Chen, D Ruan, J Xiao, L Wang, B Sun, R Saouaf, W Yang, D Li, and Z Fan. Fully automated multiorgan segmentation in abdominal magnetic resonance imaging with deep neural networks. *Medical Physics*, 2020. doi: 10.1002/mp.14429. cited By 0.
- [203] Yuankai Huo, Jiaqi Liu, Zhoubing Xu, Robert L. Harrigan, Albert Assad, Richard G. Abramson, and Bennett A. Landman. Multi-atlas segmentation enables robust multi-contrast MRI spleen segmentation for splenomegaly. In Martin A. Styner and Elsa D. Angelini, editors, *Medical Imaging 2017: Image Processing*, volume 10133, pages 72 – 79. International Society for Optics and Photonics, SPIE, 2017. doi: 10.1117/12.2254147. URL <https://doi.org/10.1117/12.2254147>.

- [204] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, M. Jorge Cardoso, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
- [205] TM MedPix. Medical image database.
- [206] Grand challenge. <https://grand-challenge.org/>.
- [207] The medical image computing and computer assisted intervention society. <http://www.miccai.org/>.
- [208] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching Wei Wang, Marc André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):1–13, 2018. ISSN 20411723. doi: 10.1038/s41467-018-07619-7.
- [209] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 12 2013. ISSN 0897-1889. doi: 10.1007/s10278-013-9622-7.
- [210] Jiang Tian, Li Liu, Zhongchao Shi, and Feiyu Xu. Automatic couinaud segmentation from ct volumes on liver using glc-unet. volume 11861 of *Lecture Notes in Computer Science*, pages 274–282. Springer International Publishing, Cham, 2019. ISBN 978-3-030-32691-3. doi: 10.1007/978-3-030-32692-0_32.
- [211] Nicholas Heller, Nikolaos Papanikolopoulos, and Christopher Weight. 2021 Kidney and Kidney Tumor Segmentation Challenge, March 2020. URL <https://doi.org/10.5281/zenodo.3714972>.
- [212] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna

- Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpal, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, pages 1–14, 2019.
- [213] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpal, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjana Sathianathan, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2020. ISSN 13618415. doi: 10.1016/j.media.2020.101821.
- [214] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L. Rubin. Ct organ segmentation using gpu data augmentation, unsupervised labels and iou loss. 11 2018.
- [215] Blaine Rister, Shivakumar Kaushik, Tomomi Nobashi, and Daniel L. Rubin. Ct-org: Ct volumes with multiple organ segmentations [dataset]. 2019. doi: 10.7937/tcia.2019.tt7f4v7o.
- [216] Ali Emre Kavur, M Alper Selver, Oğuz Dicle, Mustafa Barış, and N Sinem Gezer. Chaos - combined (ct-mr) healthy abdominal organ segmentation challenge data. 4 2019. doi: 10.5281/zenodo.3362844.
- [217] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonig, Rachana Sathish, Ronnie Rajan, Debodoot Sheet, Gurbandurdy Dvletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Chaos challenge – combined (ct-mr) healthy abdominal organ segmentation. 2020.
- [218] A. Emre Kavur, Naciye Sinem Gezer, Mustafa Barış, Yusuf Sahin, Savas Ozkan, Bora Baydar, Ulas Yuksel, Caglar Kilicier, Sahin Olut, Gozde Bozdagi Akar, Gozde Unal, Oguz Dicle, and M. Alper Selver. Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. *Diagnostic and Interventional Radiology*, 26(1):11–21, 2020. ISSN 13053612. doi: 10.5152/dir.2019.19025. cited By 4.

- [219] Holger R. Roth, Amal Farag, Evrim B. Turkbey, Le Lu Jiamin Liu, and Ronald M. Summers. Data from pancreas-ct. the cancer imaging archive. 2016. doi: 10.7937/K9/TCIA.2016.tNB1kqBU.
- [220] Holger R. Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. 6 2015.
- [221] Zhoubing Xu, Christopher P. Lee, Mattias P. Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G. Abramson, and Bennett A. Landman. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, 8 2016. ISSN 0018-9294. doi: 10.1109/TBME.2016.2574816.
- [222] BA Landman, Z Xu, JE Igelsias, M Styner, TR Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault - workshop and challenge. 2015. doi: 10.7303/syn3193805.
- [223] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2 2020. ISSN 2078-2489. doi: 10.3390/info11020125.
- [224] Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. 3 2020.
- [225] Alexander B Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. 2020. Online; accessed 01-Feb-2020.
- [226] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.09.013. cited By 174.
- [227] Davide Chicco. Siamese neural networks: An overview. pages 73–94. 2021. doi: 10.1007/978-1-0716-0826-5_3.
- [228] JANE BROMLEY, JAMES W. BENTZ, LÉON BOTTOU, ISABELLE GUYON, YANN LECUN, CLIFF MOORE, EDUARD SÄCKINGER, and ROOPAK SHAH. Signature verification using a “siamese” time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 07(04):669–688, 8 1993. ISSN 0218-0014. doi: 10.1142/s0218001493000339.

- [229] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H.S. Torr. Fully-convolutional siamese networks for object tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9914 LNCS:850–865, 2016. ISSN 16113349. doi: 10.1007/978-3-319-48881-3_56.
- [230] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. volume 2, 2015.
- [231] Kai-jian Xia, Hong-sheng Yin, and Yu-dong Zhang. Deep semantic segmentation of kidney and space-occupying lesion area based on scnn and resnet models combined with sift-flow algorithm. *Journal of Medical Systems*, 43(1):2, 1 2019. ISSN 0148-5598. doi: 10.1007/s10916-018-1116-1.
- [232] P. Lamata, A. Jalote-Parmar, F. Lamata, and J. Declerck. The resection map, a proposal for intraoperative hepatectomy guidance. *International Journal of Computer Assisted Radiology and Surgery*, 3(3-4):299–306, 9 2008. ISSN 1861-6410. doi: 10.1007/s11548-008-0226-5.
- [233] A. Radtke, S. Nadalin, G. C. Sotiropoulos, E. P. Molmenti, T. Schroeder, C. Valentin-Gamazo, H. Lang, M. Bockhorn, H.O. Peitgen, C. E. Broelsch, and M. Malagó. Computer-assisted operative planning in adult living donor liver transplantation: A new way to resolve the dilemma of the middle hepatic vein. *World Journal of Surgery*, 31(1):175–185, 1 2007. ISSN 0364-2313. doi: 10.1007/s00268-005-0718-1.
- [234] E-Liang Chen, Pau-Choo Chung, Ching-Liang Chen, Hong-Ming Tsai, and Chein-I Chang. An automatic diagnostic system for ct liver image classification. *IEEE Transactions on Biomedical Engineering*, 45(6):783–794, 6 1998. ISSN 00189294. doi: 10.1109/10.678613.
- [235] M. Gletsos, S.G. Mougiakakou, G.K. Matsopoulos, K.S. Nikita, A.S. Nikita, and D. Kelekis. A computer-aided diagnostic system to characterize ct focal liver lesions: design and optimization of a neural network classifier. *IEEE Transactions on Information Technology in Biomedicine*, 7(3):153–162, 9 2003. ISSN 1089-7771. doi: 10.1109/TITB.2003.813793.
- [236] K Yasaka, H Akai, O Abe, and S Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: A preliminary study. *Radiology*, 286(3):887–896, 2018. doi: 10.1148/radiol.2017170706. cited By 110.
- [237] C COUINAUD. [liver lobes and segments: notes on the anatomical architecture and surgery of the liver]. *La Presse medicale*, 62(33):709–712, 5 1954. ISSN 0032-7867 (Print).

- [238] Thomas S. Helling and Benoit Blondeau. Anatomic segmental resection compared to major hepatectomy in the treatment of liver neoplasms. *HPB*, 7(3):222–225, 9 2005. ISSN 1365182X. doi: 10.1080/13651820510028828.
- [239] Dário AB A.B. Oliveira, Raul Q. Feitosa, and Mauro M. Correia. Segmentation of liver, its vessels and lesions from ct images for surgical planning. *BioMedical Engineering OnLine*, 10(1):30, 2011. ISSN 1475-925X. doi: 10.1186/1475-925X-10-30.
- [240] Terry S Yoo, Michael J Ackerman, William E Lorensen, Will Schroeder, Vikram Chalana, Stephen Aylward, Dimitris Metaxas, and Ross Whitaker. Engineering and algorithm design for an image processing api: a technical report on itk—the insight toolkit. *Studies in health technology and informatics*, 85:586–592, 2002. ISSN 0926-9630 (Print).
- [241] Xiaopeng Yang, Do Jae Yang, Hong Pil Hwang, Hee Chul Yu, Sungwoo Ahn, Bong-Wan Wan Kim, and Heecheon You. Segmentation of liver and vessels from ct images and classification of liver segments for preoperative liver surgical planning in living donor liver transplantation. *Computer Methods and Programs in Biomedicine*, 158:41–52, 5 2018. ISSN 01692607. doi: 10.1016/j.cmpb.2017.12.008.
- [242] Evgin Goceri, Zarine K. Shah, and Metin N. Gurcan. Vessel segmentation from abdominal magnetic resonance images: adaptive and reconstructive approach. *International Journal for Numerical Methods in Biomedical Engineering*, 33(4):e2811, 4 2017. ISSN 20407939. doi: 10.1002/cnm.2811.
- [243] Yanling Chi, Jimin Liu, S. K. Venkatesh, Su Huang, Jiayin Zhou, Qi Tian, and W. L. Nowinski. Segmentation of liver vasculature from contrast enhanced ct images using context-based voting. *IEEE Transactions on Biomedical Engineering*, 58(8):2144–2153, 8 2011. ISSN 0018-9294. doi: 10.1109/TBME.2010.2093523.
- [244] Ye-zhan Zeng, Yu-qian Zhao, Ping Tang, Miao Liao, Yi-xiong Liang, Sheng-hui Liao, and Bei-ji Zou. Liver vessel segmentation and identification based on oriented flux symmetry and graph cuts. *Computer Methods and Programs in Biomedicine*, 150:31–39, 10 2017. ISSN 01692607. doi: 10.1016/j.cmpb.2017.07.002.
- [245] Vitoantonio Bevilacqua, Lucia Cariello, Donatello Columbo, Domenico Daleno, Massimiliano Dellisanti Fabiano, Marco Giannini, Giuseppe Mastronardi, Marcello Castellano, Massimiliano Dellisanti Fabiano, Marco Giannini, Giuseppe Mastronardi, Marcello Castellano, Massimiliano Dellisanti Fabiano, Marco Giannini, Giuseppe Mastronardi, and Marcello Castellano. Retinal fundus biometric analysis for personal identifications. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5227 LNAI, pages 1229–1237. 2008. ISBN 3540859837. doi: 10.1007/978-3-540-85984-0_147.
- [246] Titinunt Kitrungrotsakul, Xian-Hua Hua Han, Yutaro Iwamoto, Lanfen Lin, Amir Hossein Foruzan, Wei Xiong, and Yen-Wei Wei Chen. Vesselnet: A deep convolutional

- neural network with multi pathways for robust hepatic vessel segmentation. *Computerized Medical Imaging and Graphics*, 75:74–83, 7 2019. ISSN 08956111. doi: 10.1016/j.compmedimag.2019.05.002.
- [247] A Ben-Cohen, I Diamant, E Klang, M Amitai, and H Greenspan. Fully convolutional network for liver segmentation and lesions detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10008 LNCS:77–85, 2016. doi: 10.1007/978-3-319-46976-8_9. cited By 77.
- [248] Avi Ben-Cohen, Eyal Klang, Ariel Kerpel, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in ct examinations. *Neurocomputing*, 275:1585–1594, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2017.10.001. cited By 16.
- [249] Q Dou, L Yu, H Chen, Y Jin, X Yang, J Qin, and P.-A. Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41:40–54, 2017. doi: 10.1016/j.media.2017.05.001. cited By 183.
- [250] Shima Rafiei, Ebrahim Nasr-Esfahani, Kayvan Najarian, Nader Karimi, Shadrokh Samavi, and S.M. M.Reza Reza Soroushmehr. Liver segmentation in ct images using three dimensional to two dimensional fully convolutional network. pages 2067–2071. IEEE, 10 2018. ISBN 978-1-4799-7061-2. doi: 10.1109/ICIP.2018.8451238.
- [251] Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng, and Dexing Kong. Automatic 3d liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, 12(2):171–182, 2017. ISSN 18616429. doi: 10.1007/s11548-016-1467-3. cited By 88.
- [252] Peijun Hu, Fa Wu, Jialin Peng, Ping Liang, and Dexing Kong. Automatic 3d liver segmentation based on deep learning and globally optimized surface evolution. *Physics in Medicine and Biology*, 61(24):8676–8698, 2016. ISSN 13616560. doi: 10.1088/1361-6560/61/24/8676. cited By 69.
- [253] Hojin Kim, Jinhong Jung, Jieun Kim, Byungchul Cho, Jungwon Kwak, Jeong Yun Jang, Sang wook Lee, June Goo Lee, and Sang Min Yoon. Abdominal multi-organ auto-segmentation using 3d-patch-based deep convolutional neural network. *Scientific Reports*, 10(1):1–9, 2020. ISSN 20452322. doi: 10.1038/s41598-020-63285-0.
- [254] K Xia, H Yin, P Qian, Y Jiang, and S Wang. Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal ct images. *IEEE Access*, 7:96349–96358, 2019. doi: 10.1109/ACCESS.2019.2929270. cited By 42.
- [255] Dong Yang, Daguang Xu, S. Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas, and Dorin Comaniciu. Automatic liver segmentation using

- an adversarial image-to-image network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:507–515, 2017. ISSN 16113349. doi: 10.1007/978-3-319-66179-7_58.
- [256] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.
- [257] Yan Qingsen, Wang Bo, Gong Dong, Zhang Dingwen, Yang Yang, You Zheng, Zhang Yanning, and Qinfeng Shi Javen. A comprehensive ct dataset for liver computer assisted diagnosis. pages 1–13, 2021.
- [258] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [259] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [260] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.
- [261] Anjany Sekuboyina, Amirhossein Bayat, Malek E. Hussein, Maximilian Löffler, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olague, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Qingyue Wei, Kevin Brown, Matthias Wolf, Alexandre Kirszenberg, Élodie Puybareau, Alexander Valentinitich, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. 2020.
- [262] Alexandra L Williams, Aisha Al-Busaidi, Patrick J Sparrow, Judith E Adams, and Richard W Whitehouse. Under-reporting of osteoporotic vertebral fractures on computed tomography. *European journal of radiology*, 69(1):179–183, 1 2009. ISSN 1872-7727 (Electronic). doi: 10.1016/j.ejrad.2007.08.028.
- [263] Malinda Vania, Dawit Mureja, and Deukhee Lee. Automatic spine segmentation from ct images using convolutional neural network via redundant generation of class labels. *Journal of Computational Design and Engineering*, 6(2):224–232, 2019. ISSN 22885048. doi: 10.1016/j.jcde.2018.05.002.
- [264] R Korez, B Ibragimov, B Likar, F Pernuš, and T Vrtovec. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. *IEEE*

- Transactions on Medical Imaging*, 34(8):1649–1662, 2015. ISSN 1558-254X. doi: 10.1109/TMI.2015.2389334.
- [265] Jianhua Yao, Joseph E Burns, Daniel Forsberg, Alexander Seitel, Abtin Rasoulia, Purang Abolmaesumi, Kerstin Hammernik, Martin Urschler, Bulat Ibragimov, Robert Konecz, Tomaz Vrtovec, Isaac Castro-Mateos, Jose M Pozo, Alejandro F Frangi, Ronald M Summers, and Shuo Li. A multi-center milestone study of clinical vertebral ct segmentation. *Computerized Medical Imaging and Graphics*, 49:16–28, 2016. ISSN 0895-6111. doi: <https://doi.org/10.1016/j.compmedimag.2015.12.006>.
- [266] Maximilian T. Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190138.
- [267] Anjany Sekuboyina, Markus Rempfler, Alexander Valentinitzsch, Bjoern H. Menze, and Jan S. Kirschke. Labeling vertebrae with two-dimensional reformations of multidetector ct images: An adversarial approach for incorporating prior knowledge of spine anatomy. *Radiology: Artificial Intelligence*, 2(2):e190074, 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190074.
- [268] Jianhua Yao, Joseph E Burns, Hector Munoz, and Ronald M Summers. Detection of vertebral body fractures based on cortical shell unwrapping. pages 509–516, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33454-2.
- [269] Nikolas Lessmann, Bram van Ginneken, Pim A. de Jong, and Ivana Išgum. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, 53:142–155, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.02.005.
- [270] Young Jae Kim, Bilegt Ganbold, and Kwang Gi Kim. Web-based spine segmentation using deep learning in computed tomography images. *Healthcare Informatics Research*, 26(1):61–67, 2020. ISSN 2093369X. doi: 10.4258/hir.2020.26.1.61.
- [271] Syed Furqan Qadri, Danni Ai, Guoyu Hu, Mubashir Ahmad, Yong Huang, Yongtian Wang, and Jian Yang. Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in ct images. *Applied Sciences*, 9(1), 2019. ISSN 2076-3417. doi: 10.3390/app9010069.
- [272] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.
- [273] Mina Zareie, Hossein Parsaei, Saba Amiri, · Malik, Shahzad Awan, and Mohsen Ghofrani. Automatic segmentation of vertebrae in 3D CT images using adaptive fast 3D pulse coupled neural networks. *Australasian Physical & Engineering Sciences in Medicine*, 41:1009–1020, 2018. doi: 10.1007/s13246-018-0702-3. URL <https://doi.org/10.1007/s13246-018-0702-3>.

- [274] Qian Chang, Jun Shi, and Zhiheng Xiao. A new 3d segmentation algorithm based on 3d pcnn for lung ct slices. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–5, 2009. doi: 10.1109/BMEI.2009.5305554.
- [275] Hyun Jin Bae, Heejung Hyun, Younghwa Byeon, Keewon Shin, Yongwon Cho, Young Ji Song, Seong Yi, Sung Uk Kuh, Jin S. Yeom, and Namkug Kim. Fully automated 3d segmentation and separation of multiple cervical vertebrae in ct images using a 2d convolutional neural network. *Computer Methods and Programs in Biomedicine*, 184, 2020. ISSN 18727565. doi: 10.1016/j.cmpb.2019.105119.
- [276] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 5(Visigrapp):124–133, 2020. doi: 10.5220/0008975201240133.
- [277] Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentitsch, Jan S. Kirschke, and Bjoern H. Menze. Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11073 LNCS(October):649–657, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00937-3_74.
- [278] World Health Organization. Worldwide cancer data. *World Cancer Research Fund*, pages 7–12, 2018. URL <https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data%0Ainternal-pdf://0.0.1.154/worldwide-cancer-data.html>.
- [279] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2): 63, 2019.
- [280] Soumya Ghose, Arnau Oliver, Robert Martí, Xavier Lladó, Joan C Vilanova, Jordi Freixenet, Jhimli Mitra, Désiré Sidibé, and Fabrice Meriaudeau. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. *Computer methods and programs in biomedicine*, 108(1):262–287, 2012.
- [281] Konstantinos Devetzis, Francesca Kum, and Richard Popert. Recent advances in systematic and targeted prostate biopsies. *Research and Reports in Urology*, 13:799, 2021.
- [282] EJ Bass, A Pantovic, M Connor, R Gabe, AR Padhani, A Rockall, H Sokhi, H Tam, M Winkler, and HU Ahmed. A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate mri for prostate cancer in men at risk. *Prostate Cancer and Prostatic Diseases*, 24(3):596–611, 2021.
- [283] Yiqiang Zhan and Dinggang Shen. Deformable segmentation of 3-d ultrasound prostate images using statistical texture matching method. *IEEE Transactions on Medical Imaging*, 25(3):256–272, 2006.

- [284] Raman Preet Singh, Savita Gupta, and U. Rajendra Acharya. Segmentation of prostate contours for automated diagnosis using ultrasound images: A survey. *Journal of Computational Science*, 21:223–231, 2017. ISSN 1877-7503. doi: <https://doi.org/10.1016/j.jocs.2017.04.016>. URL <https://www.sciencedirect.com/science/article/pii/S1877750317304611>.
- [285] Sydney Jones and Kevin R Carter. Sonography endorectal prostate assessment, protocols, and interpretation. 2021.
- [286] Gaurav Garg and Mamta Juneja. A survey of prostate segmentation techniques in different imaging modalities. *Current Medical Imaging*, 14(1):19–46, 2018.
- [287] Ulf-Håkan Stenman, Jari Leinonen, Wan-Ming Zhang, and Patrik Finne. Prostate-specific antigen. *Seminars in Cancer Biology*, 9(2):83–93, 1999. ISSN 1044-579X. doi: <https://doi.org/10.1006/scbi.1998.0086>. URL <https://www.sciencedirect.com/science/article/pii/S1044579X98900864>.
- [288] Tristan Barrett, Arumugam Rajesh, Andrew B Rosenkrantz, Peter L Choyke, and Baris Turkbey. Pi-rads version 2.1: one small step for prostate mri. *Clinical radiology*, 74(11):841–852, 2019.
- [289] Leonard Marks, Shelena Young, and Shyam Natarajan. Mri–ultrasound fusion for guidance of targeted prostate biopsy. *Current opinion in urology*, 23(1):43, 2013.
- [290] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.
- [291] Yi Wang, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni. Deep attentional features for prostate segmentation in ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–530. Springer, 2018.
- [292] Yi Wang, Haoran Dou, Xiaowei Hu, Lei Zhu, Xin Yang, Ming Xu, Jing Qin, Pheng-Ann Heng, Tianfu Wang, and Dong Ni. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. *IEEE transactions on medical imaging*, 38(12):2768–2778, 2019.
- [293] S Sara Mahdavi, Nick Chng, Ingrid Spadinger, William J Morris, and Septimiu E Salcudean. Semi-automatic segmentation for prostate interventions. *Medical Image Analysis*, 15(2):226–237, 2011.
- [294] Lixin Gong, Sayan D Pathak, David R Haynor, Paul S Cho, and Yongmin Kim. Parametric shape modeling using deformable superellipses for prostate segmentation. *IEEE transactions on medical imaging*, 23(3):340–349, 2004.

- [295] Laurent Saroul, Olivier Bernard, Didier Vray, and Denis Friboulet. Prostate segmentation in echographic images: A variational approach using deformable super-ellipse and rayleigh distribution. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 129–132. IEEE, 2008.
- [296] Franck Grillet, Julien Behr, Paul Calame, Sébastien Aubry, and Eric Delabrousse. Acute pulmonary embolism associated with covid-19 pneumonia detected with pulmonary ct angiography. *Radiology*, 296(3):E186–E188, 2020. doi: 10.1148/radiol.2020201544. PMID: 32324103.
- [297] Ian Léonard-Lorant, Xavier Delabranche, François Séverac, Julie Helms, Coralie Pautet, Olivier Collange, Francis Schneider, Aissam Labani, Pascal Bilbault, Sébastien Molière, Pierre Leyendecker, Catherine Roy, and Mickaël Ohana. Acute pulmonary embolism in patients with covid-19 at ct angiography and relationship to d-dimer levels. *Radiology*, 296(3):E189–E191, 2020. doi: 10.1148/radiol.2020201561. PMID: 32324102.
- [298] Arnaldo Scardapane, Laura Villani, Davide Fiore Bavaro, Francesca Passerini, Amato Antonio Stabile Ianora, Nicola Maria Lucarelli, Gioacchino Angarano, Piero Portincasa, Vincenzo Ostilio Palmieri, and Annalisa Saracino. Pulmonary artery filling defects in covid-19 patients revealed using ct pulmonary angiography: A predictable complication? *BioMed Research International*, 2021, 2021. ISSN 23146141. doi: 10.1155/2021/8851736.
- [299] Davide Fiore Bavaro, Mariacristina Poliseno, Arnaldo Scardapane, Alessandra Belati, Nicolò De Gennaro, Amato Antonio Stabile Ianora, Giacchino Angarano, and Annalisa Saracino. Occurrence of acute pulmonary embolism in covid-19-a case series. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases*, 98:225–226, 9 2020. ISSN 1878-3511 (Electronic). doi: 10.1016/j.ijid.2020.06.066.
- [300] Himadri Mukherjee, Subhankar Ghosh, Ankita Dhar, Sk. Md. Obaidullah, K C Santosh, and Kaushik Roy. Shallow convolutional neural network for covid-19 outbreak screening using chest x-rays. 4 2020. doi: 10.36227/techrxiv.12156522.v1.
- [301] Rocco Cobelli, Maurizio Zompatori, Giovanni De Luca, Gianfranco Chiari, Paolo Bresciani, and Carla Marcato. Clinical usefulness of computed tomography study without contrast injection in the evaluation of acute pulmonary embolism. *Journal of Computer Assisted Tomography*, 29(1):6–12, 2005. ISSN 03638715. doi: 10.1097/01.rct.0000148274.45419.95.
- [302] Martine Remy-Jardin, Jean Baptiste Faivre, Rainer Kaergel, Antoine Hutt, Paul Felloni, Suonita Khung, Anne Laure Lejeune, Jessica Giordano, and Jacques Remy. Machine learning and deep neural network applications in the thorax: Pulmonary embolism, chronic thromboembolic pulmonary hypertension, aorta, and chronic obstructive pulmonary disease. *Journal of Thoracic Imaging*, 35(May):S40–S48, 2020. ISSN 15360237. doi: 10.1097/RTI.0000000000000492.

- [303] Hazem A Z Yousef. The accuracy of non-contrast chest computed tomographic scan in the detection of pulmonary thromboembolism. pages 61–66, 2019. doi: 10.4103/JCMRP.JCMRP.
- [304] Simon Sun, Alexandre Semionov, Xuanqian Xie, John Kosiuk, and Benoît Mesurolle. Detection of central pulmonary embolism on non-contrast computed tomography: A case control study. *International Journal of Cardiovascular Imaging*, 30(3):639–646, 2014. ISSN 15730743. doi: 10.1007/s10554-013-0356-x.
- [305] Georgia D Tourassi, Carey E Floyd, H Dirk Sostman, and R Edward Coleman. Acute pulmonary embolism: artificial neural network approach for diagnosis. *Radiology*, 189(2):555–558, 1993.
- [306] Oscar Jimenez-del Toro, Yashin Dicente Cid, Alexandra Platon, Anne Lise Hachulla, Frédéric Lador, Pierre Alexandre Poletti, and Henning Müller. A lung graph model for the radiological assessment of chronic thromboembolic pulmonary hypertension in ct. *Computers in Biology and Medicine*, 125(May):103962, 2020. ISSN 18790534. doi: 10.1016/j.compbiomed.2020.103962.
- [307] Yu Huan Wu, Shang Hua Gao, Jie Mei, Jun Xu, Deng Ping Fan, Chao Wei Zhao, and Ming Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *arXiv*, pages 1–11, 2020. ISSN 23318422.
- [308] Younes Akbari, Hanadi Hassen, Somaya Al-maadeed, and Susu Zughaier. Covid-19 lesion segmentation using lung ct scan images: Comparative study based on active contour models. pages 1–19, 2020. doi: 10.21203/rs.3.rs-40406/v1.
- [309] Yukun Cao, Zhanwei Xu, Jianjiang Feng, Cheng Jin, Xiaoyu Han, Hanping Wu, and Heshui Shi. Longitudinal assessment of covid-19 using a deep learning-based quantitative ct pipeline: Illustration of two cases. *Radiology: Cardiothoracic Imaging*, 2(2):e200082, 2020. doi: 10.1148/ryct.2020200082.
- [310] Venkatesan Rajinikanth, Seifedine Kadry, Krishnan Palani Thanaraj, Krishnamurthy Kamalanand, and Sanghyun Seo. Firefly-algorithm supported scheme to detect covid-19 lesion in lung ct scan images using shannon entropy and markov-random-field. *arXiv*, (April), 2020. ISSN 23318422.
- [311] V. Rajinikanth, Nilanjan Dey, Alex Noel Joseph Raj, Aboul Ella Hassanien, K. C. Santosh, and N. Sri Madhava Raja. Harmony-search and otsu based system for coronavirus disease (covid-19) detection using lung ct scan images. *arXiv*, (April 2020), 2020. ISSN 23318422.
- [312] Aram Ter-Sarkisov. One shot model for the prediction of covid-19 and lesions segmentation in chest ct scans through the affinity among lesion mask features. *medRxiv*, pages 2020–12, 2021.

- [313] Jinyu Zhao, Xuehai He, Xingyi Yang, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. Covid-ct-dataset: A ct image dataset about covid-19. *arXiv*, pages 1–14, 2020. ISSN 23318422.
- [314] Bo Wang, Shuo Jin, Qingsen Yan, Haibo Xu, Chuan Luo, Lai Wei, Wei Zhao, Xuexue Hou, Wenshuo Ma, Zhengqing Xu, Zhuozhao Zheng, Wenbo Sun, Lan Lan, Wei Zhang, Xiangdong Mu, Chenxi Shi, Zhongxiao Wang, Jihae Lee, Zijian Jin, Minggui Lin, Hongbo Jin, Liang Zhang, Jun Guo, Benqi Zhao, Zhizhong Ren, Shuhao Wang, Wei Xu, Xinghuan Wang, Jianming Wang, Zheng You, and Jiahong Dong. Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system. *Applied Soft Computing*, 98:106897, 2021. ISSN 15684946. doi: 10.1016/j.asoc.2020.106897.
- [315] Adel Oulefki, Sos Agaian, Thaweesak Trongtirakul, and Azzeddine Kassah Laouar. Automatic covid-19 lung infected region segmentation and measurement using ct-scans images. *Pattern Recognition*, (January):107747, 11 2020. ISSN 00313203. doi: 10.1016/j.patcog.2020.107747.
- [316] Chuansheng Zheng, Xianbo Deng, Qiang Fu, Qiang Zhou, Jiawei Feng, Hui Ma, Wenyu Liu, and Xinggong Wang. Deep learning-based detection for covid-19 from chest ct using weak label. *medRxiv*, pages 1–13, 2020. doi: 10.1101/2020.03.12.20027185.
- [317] Xinggong Wang, Xianbo Deng, Qing Fu, Qiang Zhou, Jiawei Feng, Hui Ma, Wenyu Liu, and Chuansheng Zheng. A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. *IEEE Transactions on Medical Imaging*, 39(8): 2615–2625, 2020. ISSN 1558254X. doi: 10.1109/TMI.2020.2995965.
- [318] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017. ISSN 15387445. doi: 10.1158/0008-5472.CAN-17-0339.
- [319] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*, 2016.
- [320] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019. ISSN 1470-2045. doi: [https://doi.org/10.1016/S1470-2045\(19\)30154-8](https://doi.org/10.1016/S1470-2045(19)30154-8). URL <https://www.sciencedirect.com/science/article/pii/S1470204519301548>.
- [321] Rebecca L. Siegel, Kimberly D. Miller, Ann Goding Sauer, Stacey A. Fedewa, Lynn F. Butterly, Joseph C. Anderson, Andrea Cercek, Robert A. Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(3):145–164, 2020. ISSN 0007-9235. doi: 10.3322/caac.21601.

- [322] Jinlai Gao, Zhangguo Shen, Zaixing Deng, and Lina Mei. Impact of tumor–stroma ratio on the prognosis of colorectal cancer: a systematic review. *Frontiers in oncology*, 11, 2021.
- [323] Thaína A Azevedo Tosta, Paulo Rogério de Faria, Leandro Alves Neves, and Marcelo Zanchetta do Nascimento. Computational normalization of h&e-stained histological images: Progress, challenges and future potential. *Artificial intelligence in medicine*, 95:118–132, 2019.
- [324] Chiao-Min Chen, Yao-Sian Huang, Pei-Wei Fang, Cher-Wei Liang, and Ruey-Feng Chang. A computer-aided diagnosis system for differentiation and delineation of malignant regions on whole-slide prostate histopathology image using spatial statistics and multidimensional densenet. *Medical physics*, 47(3):1021–1033, 2020.
- [325] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal, and Jeroen Van Der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 160–163. IEEE, 2017.
- [326] Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen AWM van der Laak, and Peter HN de With. Stain normalization of histopathology images using generative adversarial networks. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 573–577. IEEE, 2018.
- [327] Justin Tyler Pontalba, Thomas Gwynne-Timothy, Ephraim David, Kiran Jakate, Dimitrios Androutsos, and April Khademi. Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks. *Frontiers in bioengineering and biotechnology*, 7:300, 2019. doi: 10.3389/fbioe.2019.00300.
- [328] Marlen Runz, Daniel Rusche, Stefan Schmidt, Martin R Weihrauch, Jürgen Hesser, and Cleo-Aron Weis. Normalization of he-stained histological images using cycle consistent generative adversarial networks. *Diagnostic Pathology*, 16(1):1–10, 2021. doi: 10.1186/s13000-021-01126-y.
- [329] Zaneta Swiderska-Chadaj, Thomas de Bel, Lionel Blanchet, Alexi Baidoshvili, Dirk Vossen, Jeroen van der Laak, and Geert Litjens. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Scientific Reports*, 10(1):1–14, 2020. doi: 10.1038/s41598-020-71420-0.
- [330] Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using gan for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.
- [331] Nina Linder, Juho Konsti, Riku Turkki, Esa Rahtu, Mikael Lundin, Stig Nordling, Caj Haglund, Timo Ahonen, Matti Pietikäinen, and Johan Lundin. Identification of

- tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology*, 7(1):22, 2012. ISSN 1746-1596. doi: 10.1186/1746-1596-7-22.
- [332] Robert M Haralick, Its'hak Dinstein, and K Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973. ISSN 21682909. doi: 10.1109/TSMC.1973.4309314.
- [333] Vitoantonio Bevilacqua, Nicola Pietroleonardo, Vito Triggiani, Antonio Brunetti, Anna Maria Di Palma, Michele Rossini, and Loreto Gesualdo. An innovative neural network framework to classify blood vessels and tubules based on haralick features evaluated in histological images of kidney biopsy. *Neurocomputing*, 228:143–153, 2017. doi: 10.1016/j.neucom.2016.09.091.
- [334] Jakob Nikolas Kather, Cleo Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:1–11, 2016. ISSN 20452322. doi: 10.1038/srep27988.
- [335] Aïcha BenTaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE transactions on medical imaging*, 37(3):792–802, 2017.
- [336] Jing Ke, Yiqing Shen, Xiaoyao Liang, and Dinggang Shen. Contrastive learning based stain normalization across multiple tumor in histopathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 571–580. Springer, 2021.
- [337] Tasleem Kausar, Adeeba Kausar, Muhammad Adnan Ashraf, Muhammad Farhan Siddique, Mingjiang Wang, Muhammad Sajid, Muhammad Zeeshan Siddique, Anwar Ul Haq, and Imran Riaz. Sa-gan: Stain acclimation generative adversarial network for histopathology image analysis. *Applied Sciences*, 12(1):288, 2021.
- [338] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017. doi: 10.1109/TMI.2017.2677499.
- [339] Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. McKay, Kevan J. Salimian, Alexander Baras, and Nicholas J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3257–3267, 2020. doi: 10.1109/TMI.2019.2927182.
- [340] Zahangir Alom, Vijayan K Asari, Anil Parwani, and Tarek M Taha. Microscopic nuclei classification, segmentation, and detection with improved deep convolutional neural networks (dcnn). *Diagnostic Pathology*, 17(1):1–17, 2022.
- [341] Jie Shu, Hao Fu, Guoping Qiu, Philip Kaye, and Mohammad Ilyas. Segmenting overlapping cell nuclei in digital histopathology images. In *2013 35th Annual International*

- Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5445–5448. IEEE, 2013.
- [342] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2015.
- [343] Sudhir Sornapudi, Ronald Joe Stanley, William V Stoecker, Haidar Almubarak, Rodney Long, Sameer Antani, George Thoma, Rosemary Zuna, and Shelliane R Frazier. Deep learning nuclei detection in digitized histology images by superpixels. *Journal of pathology informatics*, 9(1):5, 2018.
- [344] Nicholas B Larson and Brooke L Fridley. Purbayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, 29(15):1888–1889, 2013.
- [345] Yinyin Yuan, Henrik Failmezger, Oscar M. Rueda, H. Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F. Schwarz, Christina Curtis, Mark J. Dunning, Helen Bardwell, Nicola Johnson, Sarah Doyle, Gulisa Turashvili, Elena Provenzano, Sam Aparicio, Carlos Caldas, and Florian Markowetz. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science Translational Medicine*, 4(157):157ra143–157ra143, 2012. doi: 10.1126/scitranslmed.3004330. URL <https://www.science.org/doi/abs/10.1126/scitranslmed.3004330>.
- [346] Manohar Kuse, Yi-Fang Wang, Vinay Kalasannavar, Michael Khan, and Nasir Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of pathology informatics*, 2(2):2, 2011.
- [347] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016. doi: 10.1109/TMI.2016.2525803.
- [348] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1):1 – 16, 2019. doi: 10.1117/1.JMI.6.1.014006. URL <https://doi.org/10.1117/1.JMI.6.1.014006>.
- [349] Daniel C. Cattran, Rosanna Coppo, H. Terence Cook, John Feehally, Ian S.D. Roberts, Stéphan Troyanov, Charles E. Alpers, Alessandro Amore, Jonathan Barratt, Francois Berthoux, Stephen Bonsib, Jan A. Bruijn, Vivette D’Agati, Giuseppe D’Amico, Steven Emancipator, Francesco Emma, Franco Ferrario, Fernando C. Fervenza, Sandrine Florquin, Agnes Fogo, Colin C. Geddes, Hermann Josef Groene, Mark Haas, Andrew M. Herzenberg, Prue A. Hill, Ronald J. Hogg, Stephen I. Hsu, J. Charles Jennette, Kensuke Joh, Bruce A. Julian, Tetsuya Kawamura, Fernand M. Lai, Chi Bon Leung, Lei Shi Li,

- Philip K.T. Li, Zhi Hong Liu, Bruce MacKinnon, Sergio Mezzano, F. Paolo Schena, Yasuhiko Tomino, Patrick D. Walker, Haiyan Wang, Jan J. Weening, Nori Yoshikawa, and Hong Zhang. The oxford classification of iga nephropathy: Rationale, clinicopathological correlations, and classification. *Kidney International*, 76:534–545, 9 2009. ISSN 00852538. doi: 10.1038/ki.2009.243.
- [350] Sean J Barbour, Gabriela Espino-Hernandez, Heather N Reich, Rosanna Coppo, Ian SD Roberts, John Feehally, Andrew M Herzenberg, Daniel C Cattran, N Bavbek, T Cook, et al. The mest score provides earlier risk prediction in lga nephropathy. *Kidney international*, 89(1):167–175, 2016. doi: 10.1038/ki.2015.322.
- [351] Shikha Purwar, Rajiv Tripathi, Adarsh Wamanrao Barwad, and AK Dinda. Detection of mesangial hypercellularity of mest-c score in immunoglobulin a-nephropathy using deep convolutional neural network. *Multimedia Tools and Applications*, 79(37):27683–27703, 2020.
- [352] Paulo Chagas, Luiz Souza, Ikaro Araújo, Nayze Aldeman, Angelo Duarte, Michele Angelo, Washington LC Dos-Santos, and Luciano Oliveira. Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artificial intelligence in medicine*, 103:101808, 2020.
- [353] Cleo-Aron Weis, Jan Niklas Bindzus, Jonas Voigt, Marlen Runz, Svetlana Hertjens, Matthias M Gaida, Zoran V Popovic, and Stefan Porubsky. Assessment of glomerular morphological patterns by deep learning algorithms. *Journal of nephrology*, pages 1–11, 2022.
- [354] Eiichiro Uchino, Kanata Suzuki, Noriaki Sato, Ryosuke Kojima, Yoshinori Tamada, Shusuke Hiragi, Hideki Yokoi, Nobuhiro Yugami, Sachiko Minamiguchi, Hironori Haga, et al. Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach. *International Journal of Medical Informatics*, 141:104231, 2020.
- [355] Noriaki Sato, Eiichiro Uchino, Ryosuke Kojima, Minoru Sakuragi, Shusuke Hiragi, Sachiko Minamiguchi, Hironori Haga, Hideki Yokoi, Motoko Yanagita, and Yasushi Okuno. Evaluation of kidney histological images using unsupervised deep learning. *Kidney international reports*, 6(9):2445–2454, 2021. doi: 10.1016/j.ekir.2021.06.008.
- [356] T Germain, S Favelier, J-P Cercueil, A Denys, D Krausé, and B Guiu. Liver segmentation: practical tips. *Diagnostic and interventional imaging*, 95(11):1003–1016, 2014.
- [357] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. volume 2017-October, pages 2999–3007. Institute of Electrical and Electronics Engineers Inc., 12 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.324.

- [358] Pei Wang and Albert C.S. Chung. Focal dice loss and image dilation for brain tumor segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:119–127, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00889-5_14.
- [359] BodyParts3D. © the database center for life science licensed under cc attribution-share alike 2.1 japan. URL <http://lifesciencedb.jp/bp3d/info/userGuide/faq/credit.html>.
- [360] Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. pages 262–270, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40763-5.
- [361] Ben Glocker, J Feulner, Antonio Criminisi, D R Haynor, and E Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. pages 590–598, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33454-2.
- [362] Judith E. Adams, Zulf Mughal, John Damilakis, and Amaka C. Offiah. *Radiology*. 2012. ISBN 9780123820402. doi: 10.1016/B978-0-12-382040-2.10012-7.
- [363] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.
- [364] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk. Nearest-neighbor methods in learning and vision. In *Neural Information Processing*, 2005.
- [365] Matthew McCormick, Xiaoxiao Liu, Luis Ibanez, Julien Jomier, and Charles Marion. Itk: enabling reproducible research and open science. *Frontiers in Neuroinformatics*, 8: 13, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00013.
- [366] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit (4th ed.)*. 2006.
- [367] G Bradski. The opencv library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- [368] Andrey Fedorov, Paul L Nguyen, Kemal Tuncali, and Clare Tempany. Annotated mri and ultrasound volume images of the prostate, 2015. URL <https://doi.org/10.5281/zenodo.16396>.
- [369] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020.
- [370] Quande Liu, Qi Dou, and Pheng Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

- [371] Antonio Brunetti, Nicola Altini, Domenico Buongiorno, Emilio Garolla, Fabio Corallo, Matteo Gravina, Vitoantonio Bevilacqua, and Berardino Prencipe. A machine learning and radiomics approach in lung cancer for predicting histological subtype. *Applied Sciences*, 12(12), 2022. ISSN 2076-3417. doi: 10.3390/app12125829.
- [372] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.
- [373] Tim McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: a survey. *Medical Image Analysis*, 1(2):91–108, 1996. ISSN 1361-8415. doi: [https://doi.org/10.1016/S1361-8415\(96\)80007-7](https://doi.org/10.1016/S1361-8415(96)80007-7). URL <https://www.sciencedirect.com/science/article/pii/S1361841596800077>.
- [374] Johan Montagnat, Hervé Delingette, and Nicholas Ayache. A review of deformable surfaces: topology, geometry and deformation. *Image and vision computing*, 19(14): 1023–1040, 2001.
- [375] Isaac Bankman. *Handbook of medical image processing and analysis*. Elsevier, 2008.
- [376] Paul J Besl. Geometric modeling and computer vision. *Proceedings of the IEEE*, 76(8):936–958, 1988.
- [377] Richard J Campbell and Patrick J Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.
- [378] Ismail B Tutar, Sayan D Pathak, Lixin Gong, Paul S Cho, Kent Wallner, and Yongmin Kim. Semiautomatic 3-d prostate segmentation from trus images using spherical harmonics. *IEEE transactions on medical imaging*, 25(12):1645–1654, 2006.
- [379] Michael Unser, Akram Aldroubi, and Murray Eden. B-spline signal processing. i. theory. *IEEE transactions on signal processing*, 41(2):821–833, 1993.
- [380] Alan H Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23, 1981.
- [381] Alex P Pentland. Perceptual organization and the representation of natural form. In *Readings in Computer Vision*, pages 680–699. Elsevier, 1987.
- [382] Alan H Barr. Global and local deformations of solid primitives. *Readings in Computer Vision*, 1(1):661–670, 1987.
- [383] Franc Solina and Ruzena Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE transactions on pattern analysis and machine intelligence*, 12(2):131–147, 1990.

- [384] S. Pieper, M. Halle, and R. Kikinis. 3d slicer. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 632–635 Vol. 1, 2004. doi: 10.1109/ISBI.2004.1398617.
- [385] Andriy Fedorov, Siavash Khallaghi, C Antonio Sánchez, Andras Lasso, Sidney Fels, Kemal Tuncali, Emily Neubauer Sugar, Tina Kapur, Chenxi Zhang, William Wells, et al. Open-source image registration for mri–trus fusion-guided prostate interventions. *International journal of computer assisted radiology and surgery*, 10(6):925–934, 2015.
- [386] Calvin R Maurer, Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, 2003.
- [387] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Josa a*, 4(4):629–642, 1987.
- [388] Daniel N Costa, Ivan Pedrosa, Francisco Donato Jr, Claus G Roehrborn, and Neil M Rofsky. Mr imaging–transrectal us fusion for targeted prostate biopsies: implications for diagnosis and clinical management. *Radiographics*, 35(3):696–708, 2015.
- [389] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Mingqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. Covid-19 ct lung and infection segmentation dataset. 4 2020. doi: 10.5281/zenodo.3757476.
- [390] Paolo Zaffino, Aldo Marzullo, Sara Moccia, Francesco Calimeri, Elena De Momi, Bernardo Bertucci, Pier Paolo Arcuri, and Maria Francesca Spadea. An open-source covid-19 ct dataset with automatic lung tissue classification for radiomics. *Bioengineering*, 8(2):1–9, 2021. ISSN 2306-5354. doi: 10.3390/bioengineering8020026.
- [391] Nguyen Quoc Khanh Le, Quang Hien Kha, Van Hiep Nguyen, Yung-Chieh Chen, Sho-Jen Cheng, Cheng-Yu Chen, et al. Machine learning-based radiomics signatures for egfr and kras mutations prediction in non-small-cell lung cancer. *International Journal of Molecular Sciences*, 22(17):9254, 2021.
- [392] Nguyen Quoc Khanh Le, Truong Nguyen Khanh Hung, Duyen Thi Do, Luu Ho Thanh Lam, Luong Huu Dang, and Tuan-Tu Huynh. Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from mri. *Computers in Biology and Medicine*, 132:104320, 2021. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.compbimed.2021.104320>. URL <https://www.sciencedirect.com/science/article/pii/S0010482521001141>.
- [393] Katja Pinker, Joanne Chin, Amy N. Melsaether, Elizabeth A. Morris, and Linda Moy. Precision medicine and radiogenomics in breast cancer: New approaches toward diagnosis and treatment. *Radiology*, 287(3):732–747, 2018. ISSN 15271315. doi: 10.1148/radiol.2018172171.

- [394] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016. ISSN 15271315. doi: 10.1148/radiol.2015151169.
- [395] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.
- [396] A Chu, Chandra M Sehgal, and James F Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11(6):415–419, 1990.
- [397] Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira, and Jean-luc Mari. Texture indexes and gray level size zone matrix application to cell nuclei classification. *Pattern Recognition and Information Processing*, (May 2014):140–145, 2009.
- [398] Chengjun Sun and William G Wee. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, 23(3):341–352, 1983. ISSN 0734-189X. doi: [https://doi.org/10.1016/0734-189X\(83\)90032-4](https://doi.org/10.1016/0734-189X(83)90032-4).
- [399] M Amadasun and R King. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5):1264–1274, 1989. doi: 10.1109/21.44046.
- [400] Fengxi Song, Zhongwei Guo, and Dayong Mei. Feature selection using principal component analysis. volume 1, pages 27–30, 2010. doi: 10.1109/ICSEM.2010.14.
- [401] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [402] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.
- [403] Michael Weiß and Markus Göker. Chapter 12 - molecular phylogenetic reconstruction. In Cletus P. Kurtzman, Jack W. Fell, and Teun Boekhout, editors, *The Yeasts (Fifth Edition)*, pages 159–174. Elsevier, London, fifth edition edition, 2011. ISBN 978-0-444-52149-1. doi: <https://doi.org/10.1016/B978-0-444-52149-1.00012-4>. URL <https://www.sciencedirect.com/science/article/pii/B9780444521491000124>.
- [404] Gregory Beylkin, Ronald Coifman, and Vladimir Rokhlin. Fast wavelet transforms and numerical algorithms. In *Fundamental Papers in Wavelet Theory*, pages 741–783. Princeton University Press, 2009.
- [405] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.
- [406] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

- [407] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.
- [408] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE transactions on Biomedical Engineering*, 61(6): 1729–1738, 2014.
- [409] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.
- [410] Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM Van Der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2015.
- [411] Nicola Altini, Tommaso Maria Marvulli, Francesco Alfredo Zito, Mariapia Caputo, Stefania Tommasi, Amalia Azzariti, Antonio Brunetti, Berardino Prencipe, Eliseo Mattioli, Simona De Summa, and Vitoantonio Bevilacqua. Colorectal cancer histology image tiles for tissue multi-class classification. 2021. doi: 10.5281/zenodo.7109754.
- [412] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. April 2018. doi: 10.5281/zenodo.1214456. URL <https://doi.org/10.5281/zenodo.1214456>.
- [413] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE, 2019.
- [414] Daniel Bug, Steffen Schneider, Anne Grote, Eva Oswald, Friedrich Feuerhake, Julia Schüler, and Dorit Merhof. Context-based normalization of histological stains using deep convolutional features. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 135–142. Springer, 2017.
- [415] Massimo Salvi, Filippo Molinari, U Rajendra Acharya, Luca Molinaro, and Kristen M Meiburger. Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification. *Computer Methods and Programs in Biomedicine Update*, 1:100004, 2021.
- [416] Yuan Chen, Yang Zhao, Wei Jia, Li Cao, and Xiaoping Liu. Adversarial-learning-based image-to-image transformation: A survey. *Neurocomputing*, 411:468–486, 2020.

- [417] Aziz Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705, 2020.
- [418] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 2021.
- [419] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [420] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020.
- [421] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.
- [422] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017.
- [423] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8983–8992, 2019.
- [424] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017.
- [425] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019.
- [426] Kutsev Bengisu Ozyoruk, Sermet Can, Guliz Irem Gokceler, Kayhan Basak, Derya Demir, Gurdeniz Serin, Uguray Payam Hacisalihoglu, Berkan Darbaz, Ming Y Lu, Tiffany Y Chen, et al. Deep learning-based frozen section to ffpe translation. *arXiv preprint arXiv:2107.11786*, 2021.
- [427] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [428] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

- [429] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [430] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [431] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.
- [432] Leonid N Wasserstein. Markov processes on countable product space describing large systems of automata. *Probl. Pered. Inform.*, 5:64–73, 1969.
- [433] Christopher X Ren, Amanda Ziemann, James Theiler, and Alice MS Durieux. Deep snow: synthesizing remote sensing imagery with generative adversarial nets. In *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXVI*, volume 11392, pages 196–205. SPIE, 2020.
- [434] Philipp Kainz, Martin Urschler, Samuel Schuler, Paul Wohlhart, and Vincent Lepetit. You should use regression to detect cells. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 276–283. Springer, 2015.
- [435] Monuseg - grand challenge. <https://monuseg.grand-challenge.org/Data/>. Accessed: April 7, 2022.
- [436] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019.
- [437] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019.
- [438] Crchistophenotypes - labeled cell nuclei data, tissue image analytics (tia) centre, warwick. https://warwick.ac.uk/fac/cross_fac/tia/data/crchistolabelednuclei. Accessed: April 7, 2022.
- [439] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019.
- [440] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

- [441] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37): eaay7120, 2019.
- [442] Anna Markella Antoniadis, Yuhang Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.
- [443] Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13943–13944, 2020.
- [444] Lixuan Du, Rongyu Zhang, and Xiaotian Wang. Overview of two-stage object detection algorithms. In *Journal of Physics: Conference Series*, volume 1544, page 012033. IOP Publishing, 2020.
- [445] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [446] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.
- [447] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.
- [448] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [449] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.