UNIVERSITÀ DEGLI STUDI DI MACERATA

Department of Humanities - Languages, Language Liaison, History, Arts, Philosophy
Ph.D. Course in Humanism and Technologies
Cycle XXXVI

Unveiling healthcare data archiving: Exploring the role of artificial intelligence in medical image analysis

Advisors
Prof. Stefano Pigliapoco
Prof. Emanuele Frontoni

Co-Advisor
Sara Moccia, PhD

Candidate
Francesca Pia Villani

Year 2024

Abstract

Medical digital archives can be seen as contemporary databases designed to store and manage vast amounts of medical information, from patient records and clinical studies to medical images and genomics data. The structured and unstructured data that compose the archives undergo rigorous curation processes, to ensure their accuracy, reliability, and standardization for clinical and research purposes.

In the rapidly evolving field of healthcare, artificial intelligence (AI) is emerging as a transformative force, able to reform medical digital archives improving the management, analysis, and retrieval of vast clinical datasets, and ultimately leading to more informed decisions, timely interventions, and improved patient outcomes. Specifically, managing medical images in digital archives poses numerous challenges such as data heterogeneity, image quality variability and lack of annotations, that can be addressed with AI solutions.

This thesis aims to exploit AI algorithms for the analysis of medical images stored in digital archives. This work investigates various medical imaging techniques, each of which is characterized by a specific application domain and consequently presents a unique set of challenges, requirements, and potential outcomes. In particular, it delves into AI diagnostic assistance for three critical imaging techniques in specific clinical scenarios:
i) Endoscopic imaging obtained during laryngoscopy examinations; this includes in-depth exploration of techniques such as keypoint detection for vocal fold motility estimation and upper aerodigestive tract cancer segmentation;
ii) Magnetic resonance imaging for intervertebral disc segmentation, for the diagnosis and treatment of spinal conditions and diseases, as well as image-guided interventions;
iii) Ultrasound imaging in rheumatology, for carpal tunnel syndrome evaluation through median nerve segmentation.

The methodologies presented in this work demonstrate the feasibility of using AI algorithms for the analysis of archived medical images, and the achieved methodological advances highlight the potential of AI-based algorithms in extracting useful information implicitly contained in digital archives.

Sommario

Gli archivi sanitari digitali possono essere considerati dei moderni database progettati per immagazzinare e gestire ingenti quantità di informazioni mediche, dalle cartelle cliniche dei pazienti, a studi clinici fino alle immagini mediche e a dati genomici. I dati strutturati e non strutturati che compongono gli archivi sanitari sono oggetto di scrupolose e rigorose procedure di validazione per garantire accuratezza, affidabilità e standardizzazione a fini clinici e di ricerca.

Nel contesto di un settore sanitario in continua e rapida evoluzione, l'intelligenza artificiale (IA) si propone come una forza trasformativa, capace di riformare gli archivi sanitari digitali migliorando la gestione, l'analisi e il recupero di vasti set di dati clinici, al fine di ottenere decisioni cliniche più informate e ripetibili, interventi tempestivi e risultati migliorati per i pazienti.

Tra i diversi dati archiviati, la gestione e l'analisi delle immagini mediche in archivi digitali presentano numerose sfide dovute all'eterogeneità dei dati, alla variabilità della qualità delle immagini, nonché alla mancanza di annotazioni. L'impiego di soluzioni basate sull'IA può aiutare a risolvere efficacemente queste problematiche, migliorando l'accuratezza dell'analisi delle immagini, standardizzando la qualità dei dati e facilitando la generazione di annotazioni dettagliate.

Questa tesi ha lo scopo di utilizzare algoritmi di IA per l'analisi di immagini mediche depositate in archivi sanitari digitali. Il presente lavoro propone di indagare varie tecniche di imaging medico, ognuna delle quali è caratterizzata da uno specifico dominio di applicazione e presenta quindi un insieme unico di sfide, requisiti e potenziali esiti. In particolare, in questo lavoro di tesi sarà oggetto di approfondimento l'assistenza diagnostica degli algoritmi di IA per tre diverse tecniche di imaging, in specifici scenari clinici:

i) Immagini endoscopiche ottenute durante esami di laringoscopia; ciò include un'esplorazione approfondita di tecniche come la detection di keypoints per la stima della motilità delle corde vocali e la segmentazione di tumori del tratto aerodigestivo superiore;

ii) Immagini di risonanza magnetica per la segmentazione dei dischi intervertebrali, per la diagnosi e il trattamento di malattie spinali, così come per lo svolgimento di interventi chirurgici guidati da immagini;

iii) Immagini ecografiche in ambito reumatologico, per la valutazione della sindrome del tunnel carpale attraverso la segmentazione del nervo mediano.

Le metodologie esposte in questo lavoro evidenziano l'efficacia degli algoritmi di IA nell'analizzare immagini mediche archiviate. I progressi metodologici ottenuti sottolineano il notevole potenziale dell'IA nel rivelare informazioni implicitamente presenti negli archivi sanitari digitali.

Contents

List of Figures vi

List of Tables ix

List of Acronyms xi

1 Introduction 1
- 1.1 Healthcare data archiving . . . 3
  - 1.1.1 Medical imaging . . . 4
- 1.2 The potential of artificial intelligence for the analysis of medical images . . . 9
  - 1.2.1 Classification . . . 10
  - 1.2.2 Segmentation . . . 10
  - 1.2.3 Detection . . . 11
  - 1.2.4 Registration . . . 11
- 1.3 Motivation and aim of the thesis . . . 12
- 1.4 Structure of the thesis . . . 13
- 1.5 Scientific publications . . . 14

2 Artificial intelligence-driven applications in endoscopic images archives **16**
- 2.1 Artificial intelligence in clinical endoscopy: insights in the field of videomics . . . 17
- 2.2 A deep learning approach for keypoints localization for vocal folds movement assessment in endoscopic images . . . 24
  - 2.2.1 Introduction . . . 25
  - 2.2.2 Materials and methods . . . 26
  - 2.2.3 Results . . . 29
  - 2.2.4 Discussion . . . 31
  - 2.2.5 Conclusion . . . 31
- 2.3 Classifying vocal folds fixation from endoscopic videos with machine learning . . . 32
  - 2.3.1 Introduction . . . 32
  - 2.3.2 Materials and methods . . . 33
  - 2.3.3 Results . . . 35
  - 2.3.4 Discussion . . . 35
  - 2.3.5 Conclusion . . . 37
- 2.4 Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes . . . 37
  - 2.4.1 Introduction . . . 38
  - 2.4.2 Materials and methods . . . 38
  - 2.4.3 Results . . . 41
  - 2.4.4 Discussion . . . 43

3 Artificial intelligence-driven applications in magnetic resonance imaging archives 47
3.1 Introduction to a self-supervised domain adaptation approach for intervertebral disc segmentation . . . 49
3.2 Materials and methods . . . 51
3.2.1 Dataset description . . . 51
3.2.2 Proposed method . . . 54
3.2.3 Pretext tasks . . . 56
3.2.4 Parameter setting . . . 56
3.2.5 Performance metrics . . . 57
3.2.6 Baseline and domain adaptation comparison . . . 57
3.2.7 Pretext task adequacy analysis . . . 58
3.3 Results . . . 58
3.4 Discussion . . . 63
3.5 Conclusion . . . 66

4 Artificial intelligence-driven applications in ultrasound images archives 67
4.1 Introduction to a deep learning approach for median nerve evaluation in ultrasound images of carpal tunnel inlet . . . 68
4.1.1 Related work . . . 69
4.2 Materials and methods . . . 71
4.2.1 Dataset . . . 72
4.2.2 Experimental setup . . . 73
4.2.3 Comparison with literature and ablation studies . . . 73
4.2.4 Performance metrics . . . 74
4.2.5 Statistical analysis . . . 75
4.3 Results . . . 75
4.4 Discussion . . . 77
4.5 Conclusion . . . 78

5 Conclusion 79
5.1 Thesis Contributions . . . 80
5.2 Impact . . . 82
5.3 Future perspectives . . . 82

Bibliography 84

List of Figures

1.1 The Picture Archiving and Communication System (PACS) workflow includes four stages: it begins with the packaging of multimodal images to Digital Imaging and Communications in Medicine (DICOM) format; the quality assurance workstation verifies patient demographics and other metadata; data is archived in a central storage device; data are retrieved in a reading Workstation, where radiologists can review the images and formulates their diagnosis. 5
1.2 Flowchart detailing the medical imaging framework in the healthcare system. At the top layer, the Electronic Health Record (EHR) system manages images and interfaces with an enterprise viewer for content viewing. The middle layer emphasizes standardized communication protocols like Digital Imaging and Communications in Medicine (DICOM) and Health Level Seven International(HL7), centered around an enterprise image repository for storage and retrieval. The bottom layer showcases diverse image sources, from native DICOM modalities to modern solutions like mobile photo apps and digital capture systems. Together, they form a comprehensive network ensuring efficient image capture, storage, and access in patient care. 6
1.3 Simplified pipeline for the creation of an artificial intelligence-based image analysis algorithm. After downloading data from the hospital's Radiology Information System (RIS)/Picture Archiving and Communication System (PACS) archive, images are deidentified and passed to a curated database for data preparation and AI algorithm development. 8

2.1 Depiction of the potential input and output of a quality assessment algorithm. 18
2.2 Example of a classification task. The algorithm distinguishes between normal and pathologic frames without identifying the area involved by the disease. 19
2.3 Image showing a bounding box localizing a laryngeal lesion. This is the typical output of detection algorithms. 20
2.4 Automatic segmentation of a laryngeal lesion provided by a Convolutional Neural Nertwork (CNN) after adequate training and optimization. 21
2.5 Endoscopic Narrow Band Imaging (NBI) frame showing an example of adjunctive data drawn from in-depth characterization by hypothetical machine learning algorithms. 23
2.6 Representation of the proposed keypoints heatmap regression network. 26
2.7 Visual sample of a labeled image. The colored points in the leftmost image represent the keypoints: left epiglottic (LE) in red, left vocal fold (LV) in yellow, anterior commissure (A) in green, right vocal fold (RV) in magenta, right epiglottic (RE) in cyan. The generated heatmaps for each keypoint are also reported separately. 27
2.8 Boxplots of the Root Mean Square Error (RMSE) for each of the five keypoints: left epiglottic (LE), left vocal fold (LV), anterior commissure (A), right vocal fold (RV), right epiglottic (RE). The proposed heatmap regression model (HR), is compared with the direct coordinate regression model (CR). Each boxplot represents the distribution of the RMSE values across all the folds. 29

2.9 Qualitative results from the comparison of the two models on sample test images. From left to right: original image, ground truth coordinate annotation, prediction of the direct coordinate regression model (CR), ground truth heatmap, predicted heatmap, prediction of the heatmap regression model (HR). 30
2.10 Representation of three consecutive frames (from left: abducted, normal, and adducted vocal cords, respectively) with ground truth keypoints annotation. The images in the first row refer to a subject with preserved motility, while the ones in the second row to a subject with fixation. The colored points represent the keypoints: left epiglottic in red, left vocal fold in yellow, anterior commissure in green, right vocal fold in magenta, right epiglottic in cyan. 33
2.11 Precision-Recall curves calculated on the test set, for all the classifiers. The best-performing classifier resulted to be XGBoost, showing the highest average precision and area under the precision-recall curve for both classes. 34
2.12 Features importance of the XGBoost classifier. Features from 0 to 8 refer to the three angles (central and externals) of the three successive frames, features from 9 to 11 refer to the static indexes of the three frames, and feature 12 refers to the dynamic index. 35
2.13 Visual samples of misclassified frames. The images in the first row were erroneously predicted as belonging to the fixation class, while the images in the second row were erroneously predicted as belonging to the preserved motility class. In the latter case, the vocal folds area occupies a small portion of the frame, which makes the prediction more challenging. 36
2.14 Visual samples of frames from the used dataset. It is characterized by high variability among the frames, which reflects also on the variability of the features used to train the models. 37
2.15 Schematic representation of the proposed architecture. The Mask R-CNN is made of a backbone (composed of a ResNet50 and a feature pyramid network), a Region Proposal Network (RPN), ROIAlign, and three heads, for classification, bounding-box regression, and segmentation. 40
2.16 Box plots detailing the diagnostic accuracy of the algorithm in different sites according to various metrics. (A) Dice Similarity Coefficient (DSC); (B) Accuracy (Acc); (C) Specificity (Spec); (D) Precision (Prec); (E) Recall (Rec); (F) Intersection over Union (IoU); (G) F1-score. 43
2.17 Visual samples of the segmentation results. From left to right: raw endoscopic frames, ground truth annotation, and predictions obtained with the proposed method. 44

3.1 The 3D intravertebral disc segmentation enables precise quantification and analysis, facilitating the evaluation of various spine-related disorders (e.g. herniated discs, degenerative disc disease, and spinal stenosis etc). 50
3.2 Intervertebral discs in magnetic resonance imaging show a wide range of variations in their appearance, including signal intensity, shape, and texture. Additionally, factors like aging, degeneration, and pathology contribute to the diverse appearances of the discs. The image showcases the distinct characteristics found within datasets, emphasizing the impact of spinal pathology on the integrity of the spine with diverse contrast levels and illuminations. Target 2, which includes images acquired with various scanners and pathological patients, exhibits additional complexity compared to the other two datasets. 52
3.3 Intensity histograms of the three datasets. The Source dataset exhibits a pronounced peak near zero, indicating a predominant concentration of low-intensity values; Target 1 presents a flatter distribution, indicating well-contrasted images; Target 2 displays a bell-shaped distribution centered around pixel intensity of 5. For visualization purposes, the plot is truncated at an intensity value of 50 and a pixel count of 250000. 52

3.4 The proposed framework for self-supervised domain adaptation which focuses on learning domain-invariant feature intensity representation. This is achieved by incorporating a pretext learning task that automatically generates labels from images of both the source and the two target domains. The pretext task and the main task, which is intervertebral disc segmentation, are simultaneously trained using multi-task learning. 53

3.5 Proposed pretext task that automatically generates labels from images of both source and target domain. 55

3.6 The qualitative results of U-Net and *t1t2s-int* training strategies on two random test images from target 1 and target 2 reveal notable differences. Upon examining the images, it becomes apparent that the U-Net approach yields sub-optimal disc segmentation, particularly for the discs located in the outermost regions of the image. 58

3.7 Boxplots of Dice Similarity Coefficient (DSC) (%) (a), Hausdorff distance (HD) (pixels) (b), Sensitivity (Sen) (%) (c), and Specificity (Spec) (%) (d) are presented for both U-Net (plum) and *t1t2s-int* (cyan) models, specifically for each intervertebral disc in T1. Intervertebral discs are named according to their anatomical position between vertebrae, where T, L, and S denote thoracic, lumbar, and sacral vertebrae respectively. 59

3.8 Boxplots of Dice Similarity Coefficient (DSC) (%) (a), Hausdorff distance (HD) (pixels) (b), Sensitivity (Sen) (%) (c), and Specificity (Spec) (%) (d) are presented for both U-Net (plum) and *t1t2s-int* (cyan) models, specifically for each intervertebral disc in T2. Intervertebral discs are named according to their anatomical position between vertebrae, where T, L, and S denote thoracic, lumbar, and sacral vertebrae respectively. 60

3.9 Qualitative results from the comparison of all the tested configurations on three random test images from each of the three datasets (source, target 1, and target 2 from top to bottom). In the source domain, the improvements brought by the proposed model (*t1t2s-int*) are particularly evident in the accurate segmentation of the contour of the discs. In target 1 and target 2, the proposed model demonstrates fewer false negatives, successfully segmenting all the discs present in the images. Yellow arrows in the image point to false positive predictions, while orange arrows to false negatives. 61

4.1 Ultrasound transverse scan sample acquired at the proximal carpal tunnel inlet. A red box includes the median nerve section; asterisks of different colors mark other relevant structures: pisiform bone profile in blue, semilunar bone profile in purple, ulnar artery in green, digital flexor tendons in yellow. 69

4.2 Schematic representation of model architecture, composed by a backbone, Region Proposal Network (RPN), and the three heads for classification, bounding-box regression and segmentation, all fed from the ROIAlign with 100 ROI candidates. The segmentation head is represented more in details as it was provided with two additional transposed layers compared with original Mask-RCNN. 71

4.3 Four visual samples of the median nerve section. From top to bottom row: original Ultrasound (US) image, ground truth mask, U-Net trained with Binary Cross Entropy (BCE) loss prediction, U-Net trained with Binary Cross Entropy-Dice Similarity Coefficient (BCE-DSC) loss prediction, Lightweight U-Net trained with BCE loss prediction, Lightweight U-Net trained with BCE-DSC loss prediction, proposed model prediction. For displaying purposes, only the upper part of the US image, which contains the median nerve section, is shown. 76

List of Tables

2.1 Results of the root mean squared error (RMSE) computed on the test sets of the five-folds. Values are expressed in pixels. 31

2.2 Performance evaluation metrics. Precision (Prec), recall (Rec), F1-score (F1), accuracy (Acc), Average Precision (AP), and Area Under the Curve (AUC) are reported. For each classifier, the first row refers to the class fixation, while the second to the class preserved motility. 36

2.3 Summary of the diagnostic performance according to the different metrics evaluated: Dice Similarity Coefficient (DSC), Accuracy (Acc), Specificity (Spec), Precision (Prec), Recall (Rec), Intersection over Union (IoU), F1-score. Values are reported as mean ± standard deviation. 42

3.1 Test sets from the three datasets. The magnetic resonance scanners used to collect the images are reported, together with the identification number (ID) of the images. S refers to the source dataset, $T1$ to the target 1 dataset, and $T2$ to the target 2 dataset. 54

3.2 Results of the performance metrics computed on the test sets of the three datasets, obtained from the baseline model (i.e. U-Net trained only on the Source dataset), the proposed model (*t1t2s-int*) and all the other dual-task models trained with various pretext task configurations. Values are reported as mean ± standard deviation. . . . 62

4.1 Overview of the Ultrasound (US) dataset characteristics in Deep Learning (DL) literature for median nerve segmentation, in terms of US acquisition site, dataset size (frames selection or frame sequences, total number of images) and patients involved in the study. 71

4.2 Performance evaluation metrics in terms of mean value and standard deviation. Mean Average Precision (mAP), Recall (Rec), Precision (Prec), and Dice Similarity Coefficient (DSC) are reported for the proposed model and the ablation studies conducted over it: Mask-R50 is the model trained using as backbone Resnet50 combined with FPN; NoAug is the model trained using no augmentations on the training data; Mask28 and Mask56 are variants of the model with a different output resolution from the segmentation head, including one and two transposed convolutional layers, respectively. 75

4.3 Comparison of segmentation performance in terms of Dice Similarity Coefficient (DSC) of the proposed model and of the U-Net and Lightweight U-Net trained using two different losses, i.e. the Binary Cross Entropy (BCE) loss and the BCE-DSC loss. 76

List of Acronyms

A Anterior commisure

Acc Accuracy

AGA Anterior Glottic Angle

AI Artificial Intelligence

AP Average Precision

AUC Area Under the Precision-Recall Curve

BCE Binary Cross-Entropy

CI Confidence Interval

CNN Convolutional Neural Network

CSA Cross-sectional Area

CT Computed Tomography

CTS Carpal Tunnel Syndrome

CW-MSE Channel-weighted Mean Square Error

DA Domain Adaptation

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DSC Dice Similarity Coefficient

EHR Electronic Health Record

EMR Electronic Medical Record

FDA Food and Drug Administration

FHIR Fast Health Interoperability Resources

FN False Negative

FP False Positive

FPN Feature Pyramid Network

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GNN Graph Neural Network

HD Hausdorff Distance

HIS Hospital Information System

HL7 Health Level Seven International

ICC Intra-class Correlation Coefficient

IHE Integrating the Healthcare Enterprise

IoU Intersection over Union

IVDs Intervertebral Disks

LE Left Epiglottic

LV Left vocal fold

MAE Mean Absolute Error

mAP Mean Average Precision

ML Machine Learning

MRI/MR Magnetic Resonance Imaging

NBI Narrow Band Imaging

NLP Natural Language Processing

PACS Picture Archiving and Communication System

PCA Principal Component Analysis

Prec Precision

RE Right Epiglottic

Rec Recall

ReLU Rectified Linear Unit

RF Random Forest

RIS Radiology Information System

RMSE Root Mean Square Error

RNN Recurrent Neural Network

ROI Region of Interest

RPN Region Proposal Network

RV	Right vocal fold
SCC	Squamous Cell Carcinoma
Sen	Sensitivity
SGD	Stochastic Gradient Descent
Spec	Specificity
SPIES	Storz Professional Image Enhancement System
SVM	Support Vector Machines
TP	True Positive
UADT	Upper Aerodigestive Tract
US	Ultrasound
VF	Vocal Folds
XGB	XGBoost

Chapter 1

Introduction

In the evolving landscape of the information era, digital archives are undergoing a profound transformation. Born out of the technological advances of digital storage in the 20th century, digital archives have evolved from virtual representations of physical repositories to living and vibrant information ecosystems. This evolution is not confined to the technological aspect; it involves a deeper transformation that has profoundly transformed the concept of data, the way users engage with it, and the paradigms governing its preservation and access. Beyond static documents, digital archives encompass a wide spectrum of multimedia and diverse digital data, meticulously curated for preservation and longevity [1]. Digital archives hold a paramount position in preserving a variety of invaluable assets: from historical and cultural preservation to research and academic preservation, to regulatory compliance [2]. Several technological milestones have been pivotal in propelling the capabilities of digital archives, and while the revolutionary impact of the Internet is undeniable, fostering global material accessibility and collaborative archival efforts [3–5], it is the integration of Artificial Intelligence (AI) that could truly amplify this impact. AI refers to the development of smart machines that mimic human cognitive functions and actions. While AI covers a broad spectrum of methods, the most significant breakthroughs have been made with Machine Learning (ML) and Deep Learning (DL). This subset of AI focuses on deriving knowledge from data without explicit programming and has propelled AI into a central technological role with the potential of revolutionizing many sectors, including healthcare. Thus, in the realm of digital archiving, AI holds the potential to improve accessibility. For instance, it can identify sensitive data, enabling archiving institutions to release non-sensitive information, or it can mark documents as pertinent to a specific search query [6, 7]. Furthermore, AI possesses the remarkable capability to harness existing corpora and labeled data, thereby automating intricate tasks such as filtering sensitive content and assisting users in navigating and interpreting vast archives. Notably, recent studies, such as those by Jaillan et al. [8], indicate an exploding interest in dialogues that explore the synergy between AI and digital archives.Traversing this dynamic landscape, the archival discourse is expanding, resonating with debates about information ownership, authenticity, and credibility, and perhaps hinting at the emergence of a new discipline defined as "Computational Archival Science" [9]. In this ever-evolving dance of technology and data, driven by the influence of AI, archives are no longer static entities, but they are unwinding towards innovative horizons of engagement, interpretation, and reimagination.

Nonetheless, despite the pervasive influence of AI across myriad sectors, its integration into archival science remains nascent. This early stage of adoption is underscored by a conspicuous lack of compelling case studies, as highlighted by Rolan [10]. Hereafter, few cases are reported.

Thibodeau [11] contends that incorporating concepts and methods from systemic-functional linguistics and graph theory can significantly enhance archival studies. In particular, systemic-functional linguistics provides insights into how language is used in human interactions, while graph theory enables the formalization and quantification of relationships between records. Notably, Graph Neural Network (GNN), a particular type of neural network designed to process data represented as graphs, and able to capture the relationships and interconnections between nodes, have been employed in clinical settings to suggest drug prescriptions based on data contained in Electronic Medical Record (EMR) [12], taking into account not only the current hospitalization records, but the entire health dossier. One of the most promising aspects of AI's role in the archival science is the automation of the archival workflow. This includes tasks such as assessment, sensitive information handling, and metadata extraction. In terms of evaluating archival records, AI can manage vast amounts of unstructured and unclassified data. Lee [13] suggests that Natural Language Processing (NLP) and ML techniques can enhance the archiving process. NLP, enabling computers to understand, interpret, and respond to human, proved to be able to identify various entities and selects the most relevant ones, while ML reduces the time needed to analyze records and automates certain classification tasks. Nevertheless, a challenge of this method is the production of training data, which can be resource-intensive, requiring thousands of manually annotated records for a ML model to properly classify a document. Recent advancements in the field of DL, such as self-supervised approaches for representation learning applied to document analysis [14, 15], aim to address the issue of manual annotation.

The handling of sensitive information in archives is one of the most challenging and urgent issues to be addressed. Often, archives, or part of their contents, are made inaccessible due to their large volume and the presence of sensitive content in the records, which currently require thorough review before any release. It is often preferable to renounce archival consultation and the resulting loss of information, rather than risk the release of sensitive information [16]. Automatic classification approaches for stored documents can be used to verify their content and estimate whether they can be released or not. Hutchinson [17, 18] proposes a supervised ML approach based on NLP and data mining for identifying sensitive content in documents containing personal data that require access restrictions.

One area where AI can be used extensively concerns archival description and metadata extraction. Automated methods can identify existing relationships between stored documents and create additional descriptive metadata to provide more context. Spencer [19] presents several techniques, including fuzzy matching, to identify different levels of similarity between stored records. Regarding metadata extraction, several studies [20–23] report the use of NLP techniques.

Another area where AI can play a role is in retrieving information from archived documents. Lee [24] uses template matching techniques and automatic classification to extract documents with the same layout from a larger corpus of documents stored in the "International Tracing Service" digital archive, one of the largest and most diverse Holocaust-related collections. Bell [25], on the other hand, takes an approach based on archival catalogs to which text mining techniques are applied, in order to automatically aggregate archival descriptions and facilitate their use.

Lastly, the increasing use of AI techniques can lead to new forms of digital archives. In fact, in addition to providing new opportunities to automate various parts of the archiving process, AI facilitates the collection of new types of data, such as from sensors or embedded systems [26].

Attention to the role of AI in archiving processes is not limited to merely providing support to the archivist's work. Recently, particular attention has been paid to the development and implementation of AI technologies, especially from the standpoint of AI ethics as a social and technical practice. Jo [27] explores the potential of archival selection and description standards, which have been developed for record-keeping, as a model for collecting training data for AI algorithms that are less biased, more transparent, and more inclusive. Finally, Mohamed [28] broadly explores decades of critical thought in the humanities and argues that this knowledge can provide a conceptual framework for the development and implementation of culturally more inclusive AI algorithms.

While the transformative potential of AI in digital archiving has been highlighted, its importance in more specialized domains, such as healthcare, is becoming increasingly evident. One such domain, healthcare, stands at the crossroads of technological advancement and the imperative to preserve critical patient data. The nuances of healthcare data archiving highlight not only the technical challenges, but also the ethical and operational considerations intrinsic to the healthcare sector. Transitioning to the next section, the intricacies and opportunities presented by AI in healthcare data archiving will be further explored, where the stakes of data preservation, accessibility, and security reach paramount importance.

1.1 Healthcare data archiving

The importance of preserving historical medical records for long-term research is not immediately obvious to everyone involved in the medical and information sectors. This raises questions such as: i) What is the state of archival practice concerning documentation of hospitals, healthcare organizations, physicians, nurses, and bio-scientists? ii) Are the problems and issues of medical archives different from those of other institutional archives, or are they the same but in a different setting?

One might wonder why it is necessary to preserve medical records that are inaccessible for decades and expensive to store and manage securely. However, a closer look at various archival medical record collections reveals their immense potential. These records provide valuable data for a wide range of medical research, and this potential is even more pronounced when considering the evolution of digital archives [29].

The advancements in digital archiving technology have had a significant impact on the way healthcare data is archived, bringing new perspectives to the practice and management of medical archives, so that all the concepts previously outlined in the context of digital archives can be translated to healthcare data archiving. Indeed, the evolution of digital archives and the technological innovations associated with it have had a profound impact on the healthcare sector.

As the medical science advances, the proliferation of EMRs and Electronic Health Record (EHR) has led to a massive increase in data. As a result, managing and storing patient information has become a major challenge today. Over the past decade, EHR adoption has grown significantly worldwide. In the United States, this growth has largely been driven by the Health Information Technology for Economic and Clinical Health (HITECH) Act, which incentivized the use of EHRs in healthcare settings. Similarly, Europe has seen progress in EHR adoption through various initiatives. The European Commission's recommendation for a European EHR exchange format aims to enhance cross-border EHR interoperability. Complementing this, the General Data Protection Regulation (GDPR) ensures the protection of personal health data. In addition, the e-Health Digital Service Infrastructure (eHDSI) and projects like the European Patient Smart Open Services (epSOS [1]) and Antilope [2] have been instrumental in developing technical specifications and frameworks for EHR interoperability and exchange across European countries. These collective efforts in the United States and Europe underscore a shared commitment to leveraging digital technology to improve healthcare outcomes, demonstrating a worldwide movement toward integrated and efficient healthcare systems through EHR adoption. EHR systems store data from each patient visit, including demographics, diagnoses, laboratory tests and outcomes, prescribed medications, medical images, and clinical observations, among others [30]. Although their primary goal is to enhance operational efficiency in healthcare, numerous researches indicate their utility in clinical informatics applications [31, 32]. Specifically, the patient information within EHRs has been leveraged for activities like extracting medical concepts, charting patient progress, deducing illnesses, and supporting clinical decision-making systems [33]. This data explosion underscores the importance of efficient patient information management, and digital archiving plays a key role in improving healthcare management. In fact, EHRs enable the efficient management of the vast amount of patient data generated daily in healthcare settings. They facilitate centralized, unified, and easily accessible storage of patient records from disparate systems, improving the process of retrieving patient information when needed. Patient-related information archived consists of 80% images and 20% text or other data. Therefore, a hybrid architecture for archiving systems is necessary. This system should use big data technologies and image compression to reduce storage requirements. It should also be reliable and ensure zero data loss [34].

As we delve deeper into the intricacies of healthcare data archiving, a relevant aspect that emerges is the management and archiving of medical images. The Picture Archiving and Communication System (PACS) plays a crucial role in modern hospitals, especially in the radiology information system, allowing for seamless management and access to medical images and patient data, and serving as a pivotal link between these two domains. Medical imaging, an essential component of modern diagnostics and patient care, generates a significant portion of the data stored in healthcare archives. This includes a variety of imaging modalities, each producing high-resolution images that require efficient storage and retrieval systems. The evolution of medical imaging technologies has not only improved diagnostic capabilities, but also increased the volume and complexity of data that must be managed. Therefore, understanding the advancements and challenges in medical imaging is crucial to understanding the full spectrum of healthcare data archiving. As we move from the general concepts of digital archiving in healthcare to the specific domain of medical imaging, we will explore targeted management and archiving techniques specifically applicable to medical images. In this context, we will also delve into the potential enhancements that AI can offer to this domain.

1.1.1 Medical imaging

Medical imaging plays an essential role in contemporary healthcare, providing clinicians with critical insights for the diagnosis, treatment, and monitoring of various diseases [35]. Preserving, protect-

---

[1] https://joinup.ec.europa.eu/collection/ehealth/solution/european-patients-smart-open-services

[2] https://cordis.europa.eu/project/id/325077

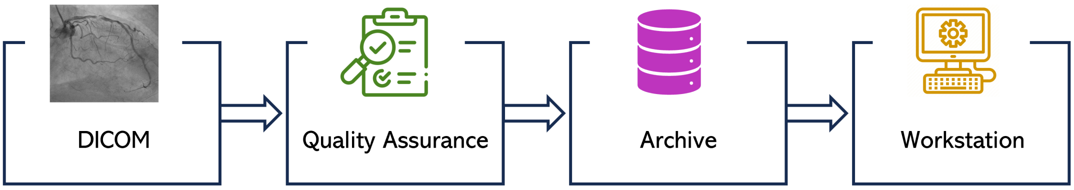


Figure 1.1: The Picture Archiving and Communication System (PACS) workflow includes four stages: it begins with the packaging of multimodal images to Digital Imaging and Communications in Medicine (DICOM) format; the quality assurance workstation verifies patient demographics and other metadata; data is archived in a central storage device; data are retrieved in a reading Workstation, where radiologists can review the images and formulates their diagnosis.

ing and archiving this data is therefore a challenge. Since the early 2000s, medical images, that were once printed on film and stored in physical film libraries, have been produced in digital format [36]. Since then, PACS has been used primarily in radiology departments to capture, store, access, view, process, and print medical images along with associated data and documents. These functionalities are typically facilitated through different degrees of local storage and server systems [36]. Nowadays, PACS can be seen as progressing within the larger framework of enterprise image management, which includes image data, metadata, medical informatics, and patient history. These developments are driven by the augmented capabilities of digital data processing, advanced databases, and fast, high-capacity networks [37]. As a direct result of these advances, image management is now comprehensive and integrates various types of medical images, including those from radiology, cardiology, surgery, endoscopy, pathology, and dermatology. This translates into tailored and optimal patient care, as physicians have access to a wider range of diagnostic tools that facilitate precise and accurate diagnosis and treatment [37].

The flow and interconnection of medical imaging and related data within a healthcare system is a complex, but crucial, orchestration that encompasses various steps and technological platforms to enhance diagnostic accuracy and patient care, thus ensuring seamless communication and management of medical information. Initially, medical images are acquired through different modalities such as Ultrasound (US), Magnetic Resonance Imaging (MRI/MR), Computed Tomography (CT), or X-ray. These images are typically sent to a PACS, which acts as a centralized storage and retrieval system. The PACS is interfaced with Radiology Information System (RIS) and Hospital Information System (HIS) to assist in patient scheduling, reporting, and billing processes. The synergy between PACS, RIS, and HIS provides a comprehensive view of patient data, enabling more informed clinical decisions. The PACS workflow (as outlined in Fig. 1.1) manages data in Digital Imaging and Communications in Medicine (DICOM) format. From a technical level, DICOM is the standard protocol used for displaying, transmitting and storing medical images. At the same time, Health Level Seven International (HL7) is a standard that facilitates the exchange, integration, and sharing of text-based medical data across different systems [38, 39]. Level Seven refers to the seventh Open Systems Interconnection (OSI) layer protocol for the health environment. Despite the widely spread of DICOM, alternative formats developed specifically for neuro-imaging are also available, such as NIfTI, MINC and ANALYZE (first version of NIfTI); more recently, the format BIDS is rapidly replacing NIfTI [40].

Recently, as we will explore in more detail later, advanced technologies such as AI, and in

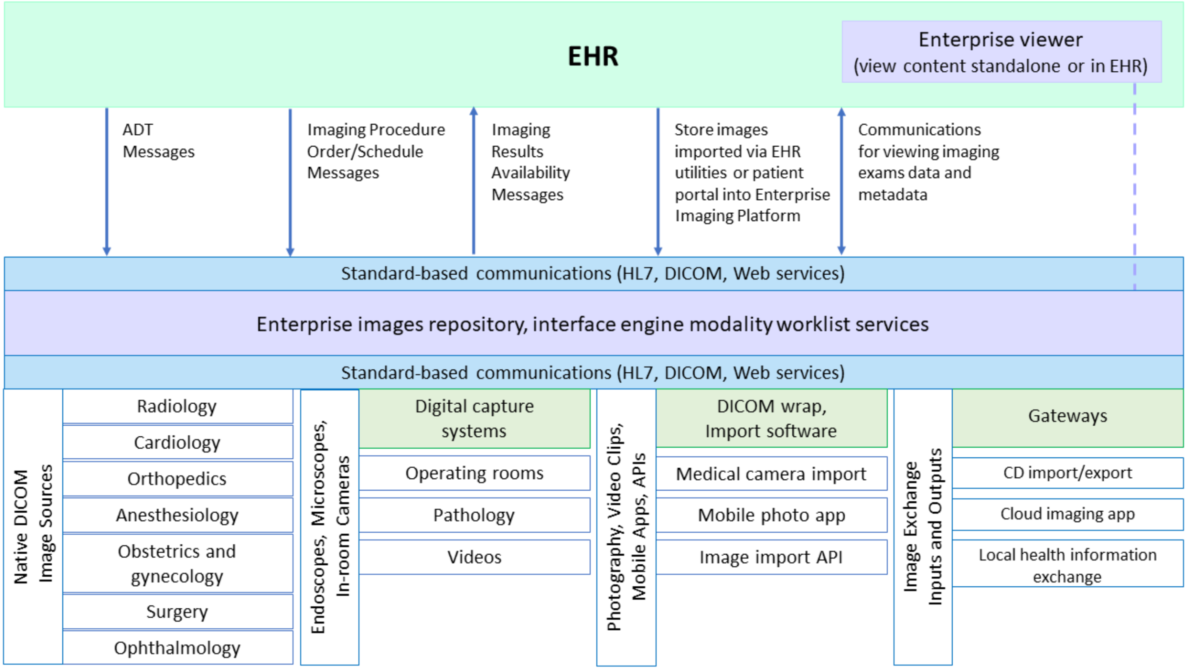

EHR
Enterprise viewer
(view content standalone or in EHR)
ADT Messages
Imaging Procedure Order/Schedule Messages
Imaging Results Availability Messages
Store images imported via EHR utilities or patient portal into Enterprise Imaging Platform
Communications for viewing imaging exams data and metadata
Standard-based communications (HL7, DICOM, Web services)
Enterprise images repository, interface engine modality worklist services
Standard-based communications (HL7, DICOM, Web services)
Native DICOM Image Sources
Radiology
Cardiology
Orthopedics
Anesthesiology
Obstetrics and gynecology
Surgery
Ophthalmology
Endoscopes, Microscopes, In-room Cameras
Digital capture systems
Operating rooms
Pathology
Videos
Photography, Video Clips, Mobile Apps, APIs
DICOM wrap, Import software
Medical camera import
Mobile photo app
Image import API
Image Exchange Inputs and Outputs
Gateways
CD import/export
Cloud imaging app
Local health information exchange

Figure 1.2: Flowchart detailing the medical imaging framework in the healthcare system. At the top layer, the Electronic Health Record (EHR) system manages images and interfaces with an enterprise viewer for content viewing. The middle layer emphasizes standardized communication protocols like Digital Imaging and Communications in Medicine (DICOM) and Health Level Seven International(HL7), centered around an enterprise image repository for storage and retrieval. The bottom layer showcases diverse image sources, from native DICOM modalities to modern solutions like mobile photo apps and digital capture systems. Together, they form a comprehensive network ensuring efficient image capture, storage, and access in patient care.

particular ML and DL, can be integrated into this ecosystem to provide enhanced image analysis, which can lead to faster and more accurate diagnoses. Health Information Exchanges (HIEs) also play a critical role in sharing medical imaging and related data among different healthcare providers and institutions, ensuring continuity of care across different healthcare settings [41]. In addition, medical imaging data can be used to enable modern healthcare approaches such as precision medicine and population health. However, due to the large size of medical imaging datasets, healthcare organizations must store them in a way that allows for efficient access by providers [42].

The workflow of medical imaging and how it interacts with various Information Technology (IT) systems in a healthcare structure is fundamental to understanding how healthcare providers manage and utilize medical imaging data to improve patient care and operational efficiency [43–45]. A schematic representation of the workflow detailing the medical imaging framework in the healthcare system is shown in Fig. 1.2.

Special mention should be made of the international standard for the exchange of medical images (i.e. DICOM) and text data (HL7). In fact, PACS and enterprise imaging systems should ensure compatibility with the currently available DICOM and HL7 standards across all vendor products. Verifying DICOM conformance involves ensuring that medical imaging or devices properly support the specific combination of functions and data types defined by the DICOM standards, which incorporate web-based services to facilitate querying (DICOM QIDO-RS), retrieving (DICOM WADO-RS) and storing (DICOM-STOW-RS) of objects such as images and medical imaging reports. This functionality enables the exchange of studies and reports as DICOM objects directly from HTML pages using the http/https protocol. The retrieved data can be either in a presentation-ready format (i.e. JPEG or GIF) or in the native DICOM format. This integration essentially bridges the gap between web technologies and medical imaging standards, providing a streamlined process for accessing and managing medical imaging data within modern web-based applications and platforms. On the other hand, through the use of HL7 Fast Health Interoperability Resources (FHIR), efforts have been made to provide patients with broader access to text and data from healthcare providers, as well as their healthcare information, both within and beyond the institutional healthcare system.

An issue related to medical images is that they are large files that require long-term archiving. Historical imaging studies must be accessible to guarantee access to the longitudinal history of patient care. In the context of the European Union, the GDPR emphasizes the protection and responsible processing of personal data, including medical images. Accordingly, these files must be stored for the shortest time necessary, determined by the purpose of data processing and legal obligations. The GDPR mandates unambiguous consent for data storage, provides for the right to data deletion, and requires data to be accurate and up-to-date. Exceptions are made for longer storage in the public interest or for scientific research, given appropriate safeguards such as anonymization. This framework ensures a balance between the confidentiality of personal health data and the necessity for access for patient care, research, and public health [46].

While the DICOM and HL7 standards provide a very technical and precise definition of the interfaces, they do not define or explain how the systems and interfaces should be used to create an interoperable environment [47]. To define how these existing standards should be used to support the workflow in the hospital, the Integrating the Healthcare Enterprise (IHE) initiative promotes the coordinated use of established standards such as DICOM and HL7. IHE helps to make clinical information systems truly interoperable [48, 49]. Therefore, large medical image archives must comply with the applicable IHE technical frameworks and integration profiles. The functionalities for managing storage within the image archive and its associated database include the following:

- Ensure that each image and data file is uniquely identified by its facility, location, and date.
- Preserve all pertinent metadata of the image that aligns with the clinical processes of its specific modality in the relevant department (, MRI in radiology or Visible Light in endoscopy).
- Adhering to relevant guidelines and laws, including those related to Health Insurance and Portability and Accountability Act (HIPAA) in the United States, and the GDPR in the European Union. This involves ensuring the security and confidentiality of personal and health data, compliance with principles such as consent, data portability, and the right to be forgotten (as stipulated by GDPR), and the secure transfer of data within, into, and out of the respective regions, as well as the duration of image study storage.

Building on this robust foundation of secure and compliant medical data management, AI integration in medical imaging stands as a potential paradigm shift in how diagnostics and patient care are approached. The ability of AI to improve the accuracy and efficiency of image analysis can enable faster, more accurate diagnoses, and personalized treatment plans. Figure 1.3 depicts the overall process of integrating AI into medical imaging, from ensuring data compliance to the development of algorithms.

The use of AI technology is contingent upon adherence to the aforementioned data protection regulations like HIPAA or GDPR, and data anonymization is central to fulfill these regulations

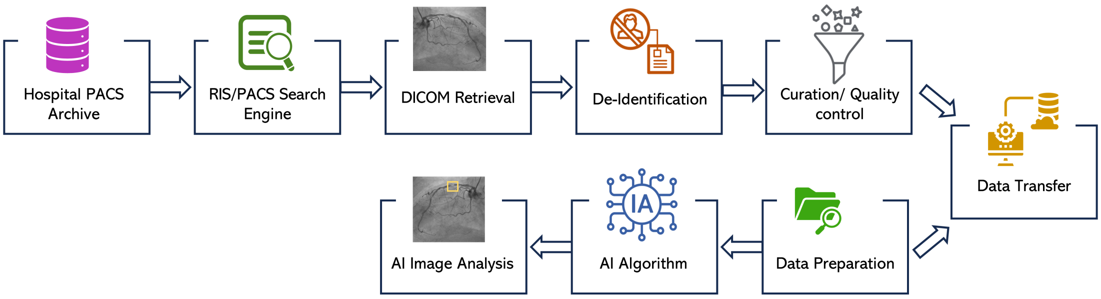


Figure 1.3: Simplified pipeline for the creation of an artificial intelligence-based image analysis algorithm. After downloading data from the hospital's Radiology Information System (RIS)/Picture Archiving and Communication System (PACS) archive, images are de-identified and passed to a curated database for data preparation and AI algorithm development.

and ensure patient privacy. Methods like de-identification, anonymization, and pseudonymization can ensure the accessibility of data while safeguarding patient confidentiality by eliminating any personally identifiable details. De-identification involves the removal or alteration of patient-specific identifiers, such as names, addresses, and hospital ID numbers, from the patient's records, for instance, within the metadata of DICOM headers. Anonymization goes a step further by erasing all patient-related data, including any additional details that could potentially reveal patient identity when combined with other available information. Pseudonymization encodes personally identifiable information using a distinct code, or pseudonym, which bears no direct link to the individual. However, if necessary, this code can be used to re-associate the data with the individual using a securely managed and isolated re-identification key [50].

Subsequently, meticulous data curation is paramount to validate the integrity of associated metadata. Data curation encompasses a set of activities post data collection — from management and standardization to validation and traceability — ensuring data quality and reproducibility. Common curation tasks involve converting, modifying, and validating DICOM files, as well as reinforcing fairness through adherence to the Findable, Accessible, Interoperable and Reusable (FAIR) Guiding Principles [51].

Furthermore, another crucial step is the medical image annotation process, which includes, among others, anatomical structures delineations and detailed lesion descriptions, playing a pivotal role not just in training AI algorithms, but also in their subsequent evaluation. Image annotation refers to the process of labelling the images with essential information (e.g. spatial location, classification), known as ground truth. This data is often contained inside the same DICOM file or in a separate text report such as JSON or CSV, which are apt for later processing and AI development. The specificity of medical image annotations (or labels) depends on the dimensional nature of the image (2D, 3D, or even 4D) and the particular imaging modality employed (e.g., MRI, US, or CT). It is important to note that while image annotation is closely related to supervised learning algorithms in AI, where the model learns from labeled data, other types of learning algorithms such as unsupervised and semi-supervised learning also play a significant role in AI development. In unsupervised learning, algorithms are trained on data without predefined labels, enabling them to discover patterns and structures within the data independently. Semi-supervised learning, on the other hand, involves a mix of labeled and unlabeled data, which can be particularly useful when there is a limited amount of labeled data available. These varying types of learning algorithms require different approaches to the use and interpretation of annotated data, thereby influencing the strategies for medical image annotation based on the intended AI application, the learning paradigm being employed, and the clinical intent behind the image acquisition [52].

In fact, medical imaging today serves a plethora of purposes, from diagnosis to treatment and follow-up, as well as from preoperative planning and intra-operative guidance to minimally invasive surgery. While AI is having a significant impact across these diverse areas, this thesis will focus specifically on its role in diagnostics.It will explore how AI is revolutionizing the way we analyze and interpret medical images for diagnostic purposes, delving into the nuances and potential of AI in enhancing diagnostic accuracy and efficiency. The broader applications of AI in medical imaging, while critical, fall outside the scope of this focused investigation.

1.2 The potential of artificial intelligence for the analysis of medical images

The advent of AI in medicine has marked a turning point in medical image analysis, a crucial aspect of modern diagnosis and health condition monitoring. DL-based techniques have been extensively investigated, with first applications appearing at workshops and conferences before being published in journals, and a noticeable increase in the number of related papers since 2015 [53]. Since then, DL has shown its great potential in the medical imaging domain, for enhancing the quality of care and improving patient outcomes. By automating medical image analysis, DL algorithms can aid in the early detection of diseases, streamline clinical workflows, and reduce the burden on healthcare professionals [53]. The incorporation of DL in medical imaging can provide specific benefits, such as a significant reduction in intra- and inter-rater variability, which is key for consistent diagnostic interpretations [54]. The ability of AI to rapidly and accurately process large image datasets can also lead to enhanced diagnostic efficiency and faster report turnaround, improving, simplifying and standardizing image acquisition, processing and reading [55]. Moreover, AI's predictive analysis capabilities in identifying risk factors and early signs of disease from medical images can contribute to proactive healthcare strategies [56]. The cost-effectiveness of AI becomes apparent through reduced need for repeat scans and optimized use of healthcare resources [57]. Furthermore, AI's adaptability allows for customized approaches in different clinical settings, providing more tailored and accurate diagnostic insights [58]. AI also serves as an invaluable tool for medical education and training, providing insights into advanced diagnostic techniques and findings [58]. A variety of DL methods have been used in medical imaging, with Convolutional Neural Network (CNN) emerging as the most common [59]. The strength of CNNs lies in their ability to autonomously learn hierarchical structures and detect local spatial features from input images, making them particularly apt for image analysis tasks [60]. While other DL methods, like Recurrent Neural Network (RNN), which are well-suited for managing sequential data, and Generative Adversarial Network (GAN), which are capable of producing new instances based on the distributions of data they have learned, have also seen applications in medical imaging [61], this work primarily focuses on methodologies based on CNNs. Therefore, discussion in this Section will center exclusively on CNNs and their applications in medical imaging.

For a thorough understanding of the role of DL in medical imaging, it is essential to examine its diverse applications. The following sections will highlight how DL techniques have been used in diverse tasks such as image classification and segmentation, as well as detection and registration.

1.2.1 Classification

Image classification was one of the first areas in which DL made a major contribution to medical image analysis. This task involves assigning a label to an input image, typically indicating the presence or absence of a particular condition or abnormality [62]. Among DL techniques, CNNs have demonstrated exceptional performance in image classification tasks [63]. In the context of medical image classification, dataset sizes are typically small compared to other computer vision applications, thus transfer learning has become particularly popular for such applications. Transfer learning refers to the use of pre-trained networks (typically on natural images) to try to solve the requirement of large datasets for deep neural network training [64]. There are two transfer learning strategies: (i) using a pre-trained network as a feature extractor, and (ii) fine-tuning a pre-trained network on medical data. In the former case, there is no need to train a network, as the extracted features are directly passed to the existing image analysis pipelines.

In the realm of deep network architectures commonly employed for classification tasks, the evolution in the medical imaging field mirrors that of computer vision applications for natural images. Initially, the focus in medical imaging was on unsupervised pre-training techniques using structures like stacked autoencoders and restricted Boltzmann machines. Stacked autoencoders are neural networks composed of multiple layers of autoencoders, where each layer learns increasingly complex data representations. Restricted Boltzmann Machines, on the other hand, are a type of stochastic neural network useful for dimensionality reduction and feature learning [65, 66]. However, recently there has been a clear shift towards the use of CNNs, noted for their wide-ranging applicability in diverse areas such as brain MRI, retinal imaging, fetal US, and lung CT scans[53, 67].

Image classification still faces significant challenges, including limited access to annotated data, disparities in class representation, and the need for model interpretability. Recently, the use of unsupervised or semi-supervised learning methods [68], as well as the improvement of data augmentation techniques [69], and the introduction of sophisticated regularization approaches [70] appear to help addressing these issues. In addition, creating techniques that offer insightful rationale for model outcomes and integrating expert domain insights into DL models can potentially boost their relevance in a clinical setting [71].

1.2.2 Segmentation

Organ and structure segmentation in medical images allows quantitative analysis of clinical parameters related to volume and shape. The segmentation task is commonly defined as the identification of the pixels (or voxels) that constitute either the contour or the interior of the objects of interest. Being the most common subject of DL application to medical images [53], segmentation had the greatest variety of methodologies, including CNNs and RNNs. The most well-known CNN architecture for medical image segmentation is U-Net [72], which combines the same number of upsampling and downsampling layers and connects them with so-called skip connections. This approach merges features from both the contracting and expanding paths and, in terms of training, it implies that U-Net can handle entire images or scans in a single forward pass, directly producing a segmentation map. This capability enables U-net to consider the complete image context, offering a potential benefit over patch-based CNNs. A 3D-variant of U-Net architecture also exists and it is known as V-Net [73], it performs 3D image segmentation using 3D convolutional layers. RNNs have recently become more popular for segmentation tasks [74], as well as fully CNN (fCNNs) applied both alone [75] or combined with graphical models to refine segmentation outputs[76].

Despite the success of segmentation models, challenges related to the need for large annotated datasets, model interpretability, and algorithms' robustness to variations in image quality, acquisition protocols, and patient populations still need to be addressed [77].

1.2.3 Detection

The localization of anatomical objects, such as organs or landmarks, results an important pre-processing step in segmentation tasks or in the clinical workflow for therapy planning and intervention. Localization in medical imaging often requires parsing of 3D volumes, and both pre-trained CNN architectures and restricted Boltzmann machines, have been used for this purpose [78, 79]. These studies approach the localization task as a classification task, thus generic DL architectures can be used. Another approach relies on directly regression of landmark locations with CNNs [80]. This method is typically addressed using landmark maps, where each landmark is represented by a Gaussian heatmap as ground truth input data, and the network is directly trained to predict this landmark map. CNNs have also been used for the localization of scan planes or key frames in temporal data [81]. The latter have also been exploited using RNNs to leverage temporal information contained in medical videos [82]. For medical images, the prevailing method for identifying organs, areas, and landmarks has been through 2D image classification using CNNs. Nonetheless, this concept has recently been expanded by modifying the learning process to directly emphasize accurate localization, with promising results [53].

On the other hand, the detection of objects of interest or lesions in medical images is one of the most labor-intensive tasks from the clinician's point of view. In fact, it consists of localizing and identifying small lesions in the entire image space. Surprisingly, the first object detection system using a four-layer CNN was proposed in 1995 to detect nodules in X-ray images [83]. Mostf DL-based object detection methods still rely on CNNs for pixel (or voxel) classification, typically followed by a post-processing step to derive potential object candidates. Since the classification task performed at each pixel is essentially object classification, the structure and techniques of CNNs closely mirror those used in classification tasks. Also in this case, as the annotation burden to generate training data is significant, weakly supervised DL approaches have been explored [84].

Class imbalance/hard-negative mining, as well as efficient pixel/voxel-wise processing of images, are challenges that still need to be addressed.

1.2.4 Registration

Registration (i.e. spatial alignment) of medical images is a common task in which a coordinate transformation is computed from one image to another. Thus, image registration involves aligning two or more images, typically acquired from different modalities or at different times, to simplify comparison and analysis [60]. There is a growing trend to employ DL in image registration, with architectures like CNNs and spatial transformer networks being predominant. For supervised learning approaches, deformation fields or similarity metrics have been adopted as labels[85]; however, given the difficulties in obtaining labeled data for registration, unsupervised learning methods, that do not require ground truth correspondences, are often adopted [86].

As for the other tasks, image registration presents challenges related to the need for large, diverse training datasets, the limited interpretability of the learned transformations, and the potential for overfitting or generating implausible deformations [60]. A potential benefit can be retrieved from the incorporation of domain knowledge into the models, or by designing robust evaluation metrics to highlight the clinical relevance of the registration results [86].

1.3 Motivation and aim of the thesis

Considering the growing importance of medical digital archives as a rich source of patient health data, and the role of medical images as a crucial component of these archives, this thesis proposes to harness the power of DL algorithms for medical image analysis. The primary focus is on the development of AI applications specifically designed to aid in the diagnostic process. By leveraging advanced DL techniques, the goal is to enhance the precision and efficiency of the diagnosis from medical images, improving clinical decision-making and patient outcomes. These archives, which have accumulated vast amounts of patient images over the years, offer a unique opportunity to glean insights into various health conditions, disease progressions, and treatment outcomes. By analyzing these images, clinical decision-making can be enhanced, and patterns previously unnoticed by human experts can be potentially uncovered.

This work explores different DL approaches tailored to improve diagnostic processes across various medical fields and performing a range of tasks, highlighting the versatility and effectiveness of DL. In particular, it concentrates on three specific imaging modalities in different clinical scenarios: i) endoscopic imaging, a field in which DL has the potential to enhance diagnostic accuracy while reducing subjectivity in frame evaluation; ii) MR imaging, in which DL techniques have the potential to automate complex tasks to analyze human anatomy and pathology; iii) US rheumatological imaging, a relatively nascent area for DL applications, which presents a plethora of untapped opportunities and challenges concerning DL methodologies.

By analyzing the current technical challenges in these domains and identifying opportunities in the less explored areas, the guiding research hypotheses for this PhD work can be summarized as:

- The vast amount of historical images available in medical archives can be effectively leveraged by DL algorithms, which have the potential to significantly enhance the analysis of current medical imaging.

- DL approaches that venture into unsolved challenges can propel both academic research and clinical practice forward. By uncovering patterns and insights from medical digital archives, they can greatly aid in improving patient care and outcomes.

- More consistent and objective diagnoses can be achieved by effectively minimizing inter- and intra-observer variability, by employing DL algorithms for the automated analysis of archived images.

The methodology crafted to delve into these hypotheses will be elaborated upon in this PhD dissertation. Furthermore, sample applications will be showcased, offering a practical lens to experimentally test the hypotheses and underline the significance of DL in making the most of medical digital archives. The development of the applications discussed in this dissertation was conducted without accessing comprehensive medical archives or PACS, primarily owing to privacy and access restrictions. Consequently, the research concentrated on utilizing datasets that were either publicly available through the scientific community or supplied by local hospitals over approval from ethical committees at the clinical sites. Despite the reduced data availability and the privacy constraints, it was possible to develop and also effectively showcase the capabilities of these algorithms on medical image domain, illustrating their significant potential in the comprehensive image archive.

1.4 Structure of the thesis

The subsequent parts of the dissertation are organized as follows:

Chapter 2 presents the crucial clinical relevance of making an accurate and quantitative diagnosis of the upper aerodigestive tract diseases. Within the chapter, the challenges behind the analysis of endoscopic images will be analyzed and a number of AI methodologies will be proposed to gradually meet the actual clinical needs. All the presented methods have a common root: supporting the clinical diagnosis of pathologies related to the larynx. In detail, vocal folds motility assessment and cancer segmentation will be addressed applying DL and ML approaches.

Chapter 3 delineates a DL-driven application for MRI archives. In particular, a new self-supervised domain approach for intervertebral disc segmentation is proposed by taking advantage of three publicly available datasets of the field. This approach holds its relevance in various medical applications associated with the spine, encompassing the diagnosis and treatment of spinal conditions and diseases, as well as image-guided interventions.

Chapter 4 aims to present an innovative quantitative assessment system for carpal tunnel syndrome. Indeed, despite its relevance, the diagnosis of this syndrome still relies on clinical history and physical examination, sometimes integrated with electrodiagnostic tests. To solve the need for a more quantitative and objective diagnosis, this chapter describes a DL method for the median nerve evaluation in US rheumatological images, paving the way for future research in the field.

Chapter 5 offers an overview of the conclusions, scientific and clinical implications of this PhD work are reported and discussed. Then, final considerations, open challenges, and future perspectives of the healthcare ecosystem are discussed.

Chapters 2 ÷ 4, which differ for the clinical need to be solved, i) give the reader an overview of the state of the art in the field; ii) present the adopted dataset; iii) justify the choice of the proposed AI pipelines; iv) present the experimental setup and evaluation metrics; v) provide the results for evaluating the performance of the proposed method; vi) discuss the obtained results highlighting the limitations, and vii) conclude with the future perspective of the research.

1.5 Scientific publications

Part of the methodologies presented and tested in this PhD dissertation were presented in the following peer-reviewed publications.

Journal publications

A. Paderno, F. Gennarini, A. Sordi, C. Montenegro, D. Lancini, **F.P. Villani**, S. Moccia, C. Piazza. *Artificial intelligence in clinical endoscopy: Insights in the field of videomics.* Frontiers in Surgery, 9, 933297, 2022, Frontiers.

A. Paderno, **F.P. Villani**, M. Fior, G. Berretti, F. Gennarini, G. Zigliani, E. Ulaj, C. Montenegro, A. Sordi, C. Sampieri. *Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes.* Acta Otorhinolaryngologica Italica, 43, 4, 283, 2023, Pacini Editore.

M.C. Fiorentino & **F.P. Villani**, R. Benito Herce, M.A. Gonzalez Ballester, A. Mancini, K. Lopez-Linares Roman. *Self-supervised Domain Adaptation for Intervertebral Disc Segmentation in Magnetic Resonance Imaging.* Currently under review at International Journal of Computer Assisted Radiology and Surgery.

M. Di Cosmo, M.C. Fiorentino; **F.P. Villani**, E. Frontoni, G. Smerilli, E. Filippucci, S. Moccia. *A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet.* Medical & Biological Engineering & Computing, 60, 11, 3255-3264, 2022, Springer Berlin Heidelberg

Conference proceedings

F.P. Villani, A. Paderno, M.C. Fiorentino, A. Casella, C. Piazza, S. Moccia. *Classifying Vocal Folds Fixation from Endoscopic Videos with Machine Learning.* 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE.

The following publications, which are only partially related to the topic of the doctorate and will not be discussed in the thesis, result from intra- and inter-research group collaborations:

L. Serrador, **F.P. Villani**, S. Moccia, C. Santos. *Knowledge distillation on individual vertebrae segmentation exploiting 3D U-Net.* Currently under review at Computerized Medical Imaging and Graphics.

G. Migliorelli, M.C. Fiorentino, M. Di Cosmo, **F.P. Villani**, A. Mancini, S. Moccia. *On the Use of Contrastive Learning for Standard-Plane Classification in Fetal Ultrasound Imaging.* Currently under review at Computers in Biology and Medicine.

M.C. Fiorentino, **F.P. Villani**, M. Di Cosmo, E. Frontoni, S. Moccia. *A review on deep-learning algorithms for fetal ultrasound-image analysis.* Medical Image Analysis, 83, 102629, 2023, Elsevier.

A. Paderno, **F.P. Villani**, A. Sordi, C. Montenegro, S. Moccia. *Deep learning in endoscopy: the importance of standardization.* ACTA Otorhinolaryngologica Italica, 1-3, 2023

M. Di Cosmo, M.C. Fiorentino, **F.P. Villani**, G. Sartini, G. Smerilli, E. Filippucci, E. Frontoni, S. Moccia. *Learning-based median nerve segmentation from ultrasound images for carpal tunnel syndrome evaluation.* 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 3025-3028, 2021, IEEE.

F.P. Villani, M. Di Cosmo, A. Bertelsen Simonetti, E. Frontoni, S. Moccia. *Development of an Augmented Reality system based on marker tracking for robotic-assisted minimally invasive spine surgery.* International Conference on Pattern Recognition, 461-475, 2021, Springer International Publishing Cham.

This thesis had also technology transfer implications. The author had the opportunity to join the department of Digital Health & Biomedical Technologies at Vicomtech, Spain. Vicomtech[3] is a technological center set up as a private non-profit Foundation. It responds to Applied Research, Development and Innovation in Information Technology, especially the convergence of Computer Graphics and Computer Vision (Visual Computing), Data Analytics and Intelligence, Interactive Digital Media and Language Technologies, in businesses and institutions in the biomedical domain. During the months spent there, the author had the possibility to focus on implementing and validating AI-based solutions for MRI.

---

[3] https://www.vicomtech.org/

Chapter 2

Artificial intelligence-driven applications in endoscopic images archives

Endoscopic imaging is a minimally invasive medical technique that employs specialized instruments equipped with integrated cameras for the inspection of the inner cavities and organs of the human body [87]. Endoscopes, which can be either rigid or flexible tubular devices, allow direct visualization within the body. These endoscopic systems can be optical, using lenses, transparent rods or fibers, and can also include integrated or add-on cameras. The endoscope's tip is inserted through small incisions or natural body orifices, with illumination provided by a light source. When cameras are employed, the field of view is displayed on a screen and can be recorded for later diagnosis or documentation. Different types of endoscopic systems are designed for specific examinations, such as those of the oral cavity, joints, lungs, abdomen, bladder or colon. In addition to diagnosis, endoscopes play an important role in assisting therapeutic procedures, including surgery and minimally invasive interventions [87].

Endoscopy finds application in a variety of medical contexts, including gastroenterology for gastrointestinal tract investigation, pulmonology for respiratory system assessment, and otolaryngology for laryngeal and vocal cord examinations. Endoscopy provides real-time, high-resolution imaging essential for diagnostic, surgical and therapeutic purposes. Nevertheless, the interpretation of endoscopic images can be challenging and often relies on the clinician's expertise.

In this field, ML and DL techniques have been developed to recognize disease patterns and predict specific characteristics that can aid clinicians in diagnosis, treatment planning, and post-treatment follow-up [88]. The analysis of endoscopic images using computer vision methods, defined as videomics, has recently focused on five broad tasks of increasing complexity: quality assessment of endoscopic images, classification of pathological and non-pathological frames, detection of lesions within frames, segmentation of pathological lesions, and in-depth characterization of neoplastic lesions [88].

Currently, DL is predominantly utilized in gastrointestinal endoscopy, due to the extensive availability of gastrointestinal endoscopy image databases. Annotated image repositories [89–91] serve as the basis for training DL algorithms, preserving the collective knowledge of the medical community and enabling improved cancer detection, diagnosis of infections, and identification of bleedings or polyps [92]. In addition to supporting less-experienced physicians, automating image analysis and emulating tasks already mastered by physicians [93], DL endeavors to excel in areas where medical professionals have had limited success. An active research focuses on surgical procedure enhancement, aiming at improving the quality and value of interventional healthcare by capturing, organizing, analyzing, and modeling data [94].

However, in most medical fields, DL for endoscopic images is still at an early stage of development, mainly due to the paucity of annotated datasets which hinders the development of robust and automatic algorithms that can be translated in clinical practice. This is particularly true for endoscopic applications that examine the oral cavity, nasal passages, oropharynx, and larynx, which are part of the field of otolaryngology and head and neck surgery.

Given the current gap in research and the limited application of DL for endoscopic procedures in the Upper Aerodigestive Tract (UADT), as highlighted by the paucity of specialized datasets and the nascent stage of algorithm development, this Chapter is motivated by the need to address these deficiencies. It aims to bridge this gap, focusing on the innovative use of DL in enhancing clinical endoscopy within otolaryngology, particularly in the assessment of Vocal Folds (VF) movement and the early detection of UADT cancers. Thus, after an in-depth examination of the role of DL in clinical endoscopy, two specific applications of DL in laryngoscopy are presented: the first explores an innovative DL approach for VF movement assessment; while the second regards a DL technique for the early detection of UADT cancer. From a clinical standpoint, the suggested frameworks demonstrated their efficacy as supportive tools in clinical practice, enhancing the reproducibility of VF motility estimation and enhancing cancer segmentation.

2.1 Artificial intelligence in clinical endoscopy: insights in the field of videomics

Diagnostic endoscopy is an essential component in the assessment of the UADT and is a cornerstone as a first-line diagnostic tool, especially after the introduction of the “bioendoscopy” concept [95]. The introduction of videoendoscopy significantly improved this field by the development of high-quality video recording, image magnification, high-definition visualization, and advanced optical filters such as Narrow Band Imaging (NBI), Storz Professional Image Enhancement System (SPIES) or Image 1S, and I-Scan. These nuances, together with the constant advancement in ML, have opened new possibilities for image analysis in a computer vision-oriented approach. Here, DL, is playing a paramount role.

In the field of supervised learning, when provided with both the “problem” (i.e., unlabeled videoendoscopic frame) and the “solution” (i.e., annotated frame or “ground truth”), DL algorithms iteratively learn their internal parameters (i.e., weights) to progressively improve diagnostic performance and specialize on a given objective. In this field, recent studies have focused on five broad tasks with increasing complexity and computational load, which can be summarized as follows:

- Quality assessment of endoscopic images;
- Classification of pathologic and non-pathologic frames;
- Detection of lesions inside frames;
- Segmentation of pathologic lesions;

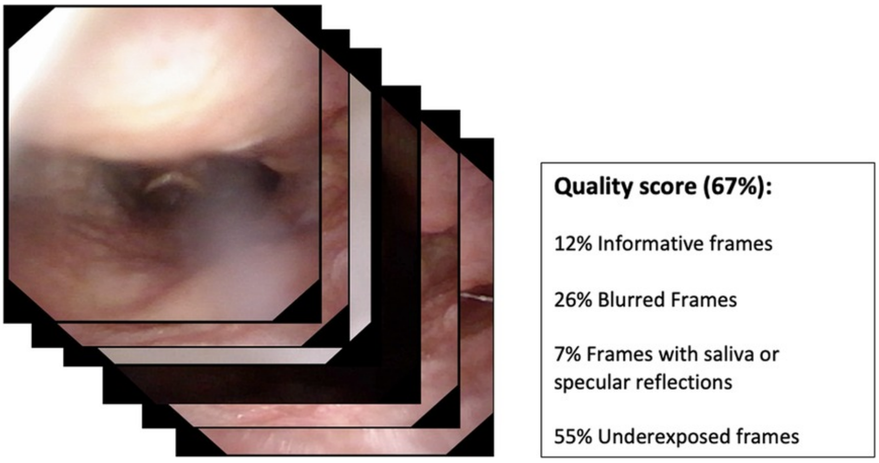


Figure 2.1: Depiction of the potential input and output of a quality assessment algorithm.

- In-depth characterization of neoplastic lesions.

In this area, a stepwise approach has the potential to make use of incremental refinements of algorithms and develop functional "minimal viable products" that can be introduced in clinical practice as early as possible, even without the full suite of the above-mentioned applications. This is especially true considering that, as mentioned, the main limiting factor in this field is the paucity of large dedicated datasets that are usable for training. Gómez et al. [96] initially addressed this issue in the field of high-speed laryngeal videoendoscopy by collecting and publishing the Benchmark for Automatic Glottis Segmentation (BAGLS), a multihospital glottis segmentation dataset. However, with the progressive expansion of the available training images, it will be possible to tackle increasingly complex challenges.

Furthermore, the application of transfer learning techniques may significantly improve algorithm training and reduce the number of images needed to achieve optimal performance. Pretraining with endoscopic images from a different anatomic site may provide an adjunctive advantage, especially in small datasets.

A potential approach to address the low number of manually annotated images is offered by unsupervised and self-supervised learning. Unlike supervised learning, which is biased toward how it is being supervised, unsupervised learning derives insights directly from the data itself, groups the data, and helps make data-driven decisions without external biases [97]. This approach may be particularly useful to cluster endoscopic frames into different categories (e.g., low visibility vs. good visibility) to help the clinician's assessment. On the other hand, self-supervised learning takes advantage of unlabeled images of the same pathology but captured from different views to significantly enhance the performance of pretraining. However, these options still need to be fully explored in the field of UADT endoscopy [98].

Quality assessment The first area in which AI can be effectively applied to diagnostic videoendoscopies is their quality control. In fact, in every examination, the majority of videoendoscopic frames are not diagnostic due to the presence of technical or patient-related factors that limit visualization. These factors, in the field of UADT evaluation, are mainly represented by repeated swallowing, gag reflex, secretions, blurring of the camera, specular reflections, and over- and underexposure. Automatic identification and classification of these issues can be of help in real-time

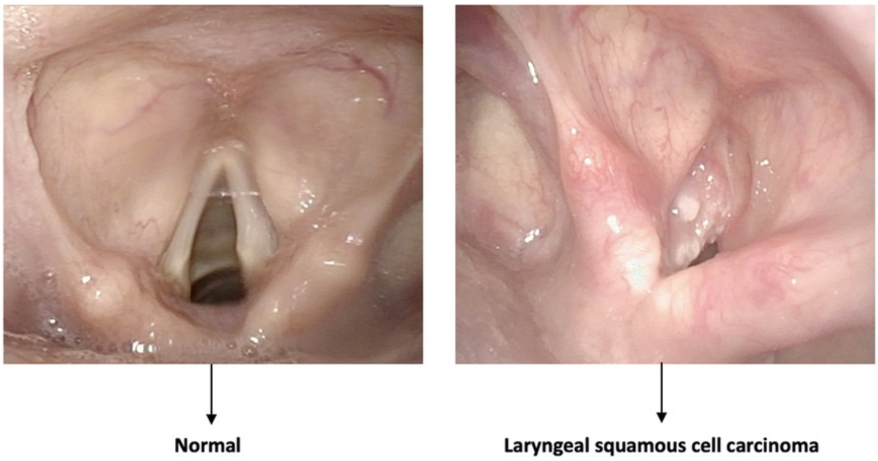


Figure 2.2: Example of a classification task. The algorithm distinguishes between normal and pathologic frames without identifying the area involved by the disease.

determination of the quality of an endoscopic examination and as depicted in Fig. 2.1 it may allow automatic detection of the most significant frames in a given recording.

In this field, Patrini et al. [99] developed a ML-based strategy for automatic selection of informative videolaryngoscopic frames. This approach resulted in a Recall (Rec) of 0.97 when classifying informative vs. uninformative frames (i.e., blurred, with saliva or specular reflections, and underexposed) with Support Vector Machines (SVM) (i.e., conventional ML algorithms) and of 0.98 with a CNN-based classification. Furthermore, their work demonstrated the potential of transfer learning in medical image analysis.

As a proof of concept, recent advances in the field of gastrointestinal endoscopy have led to the development of a fully automatic framework that can detect and classify different artifacts, segment artifact instances, provide a quality score for each frame, and restore partially corrupted frames [100].

Classification Classification is a typical DL task, and in the field of videoendoscopy can be applied to distinguish between normal and pathological mucosa, as illustrated in Fig. 2.2.

In this field, He et al. [101] applied CNN to interpret images of laryngeal squamous cell carcinoma using static NBI frames to determine whether a lesion was benign or malignant. The model reached an Accuracy (Acc) of 90.6%, a Sensitivity (Sen) of 88.8%, and a Specificity (Spec) of 92.2%. Furthermore, the authors demonstrated that the model accuracy in distinguishing malignant lesions was higher than that of human experts. A similar approach was described by Esmaeili et al. [102], training a CNN for the automatic classification of NBI images into benign and malignant. A pretrained ResNet50 architecture was adopted, and three experiments with several models were generated and validated. The model showed a striking diagnostic performance and achieved a testing Acc of 0.83.

Considering multiple classification groups, Zhao et al. [103] proposed a four class-system of vocal cord targets (i.e., normal mucosa, polyp, keratinization, and carcinoma), and a laryngoscopy dataset was divided into “urgent” (keratinization, carcinoma) and “nonurgent” (normal mucosa, polyp) cases. An overall Acc of 80.2%, an F1-score of 0.78, and an Area Under the Precision-Recall Curve (AUC) of 0.96 were achieved. The proposed method delivered high classification performance

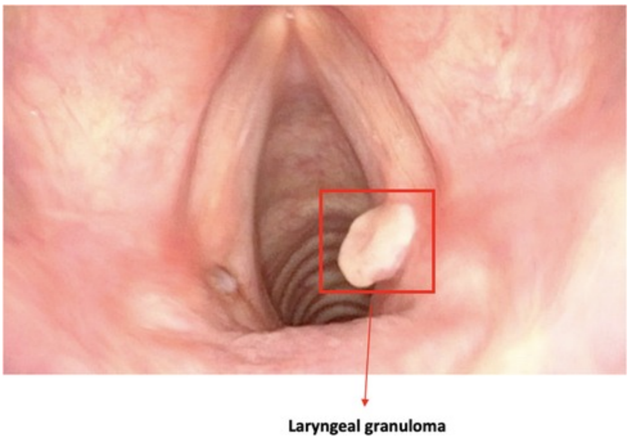


Figure 2.3: Image showing a bounding box localizing a laryngeal lesion. This is the typical output of detection algorithms.

of normal mucosa, polyps, and carcinoma in an extremely quick time.

Other studies [104, 105] have employed ML to classify pharyngo-laryngeal benign lesions during videoendoscopy, demonstrating notable results. A preliminary attempt was described in 2014 by Huang et al. [104], who proposed an automatic system aimed at recognizing the dynamic image of the glottis and classifying different VF disorders ("normal VF," "VF paralysis," "VF polyp," and "VF cyst"). This study used an SVM classifier and reached an Acc of 98.7%. However, the patterns to be classified did not include dysplasia or malignancy. Following [106], Dunham et al. [105] proposed the concept of "optical biopsy" using CNNs. The first objective was to classify endoscopic images into one of five benign classes (normal mucosa, nodules, papilloma, polyps, and webs). The second was, using a binary classifier, to distinguish malignant/premalignant from benign lesions. The overall Acc for the multiclass benign VF lesion classifier was 80.8%, while the binary test achieved an overall Acc of 93%.

Different authors [107, 108] also demonstrated the feasibility of classifying oropharyngeal and oral cavity lesions using ML technology. For the oropharynx, Mascharak et al. [107] used a naive Bayesian classifier (color and texture) to demonstrate the value of NBI imaging instead of white light videoendoscopy, which added more definition to tumor margins and highlighted submucosal vascularization. Five-fold cross-validation provided an AUC of over 80% for NBI and under 55% for white light endoscopy models ($p < 0.001$).

In the oral cavity, in 2018, Song et al. [108], employing CNNs, proposed a low-cost, smartphone-based, automatic image classification system. The authors collected data from 190 patients across several centers in India to detect oral dysplasia and malignancy using a dual-mode image analysis with white light and autofluorescence. The study compared the accuracy of the single- (white light or autofluorescence) and dual-mode (white light and autofluorescence) image analysis, demonstrating that the latter had a better diagnostic performance. The final model reached an Acc of 87%, a Sen of 85%, and a Spec of 89%.

Detection Lesion detection (Fig. 2.3) remains the main objective of DL-based strategies in contemporary clinical videoendoscopy. Different authors have described the potential of CNN in the detection of cancer, premalignant lesions, benign lesions, and normal tissue.

Inaba et al. [109] trained a CNN-based algorithm (RetinaNet) to detect superficial laryngo-

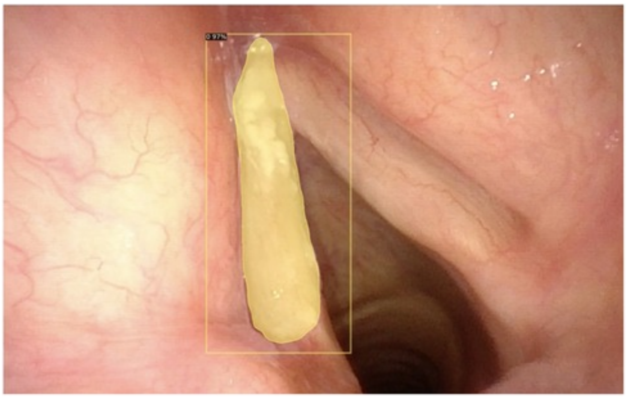


Figure 2.4: Automatic segmentation of a laryngeal lesion provided by a Convolutional Neural Nertwork (CNN) after adequate training and optimization.

pharyngeal cancer. To evaluate diagnostic accuracy, 400 pathologic images and 800 of normal mucosa were collected, reaching an Acc, Sen, and Spec of 97%, 95%, and 98%, respectively. The definition of correct diagnosis was set with an Intersection over Union (IoU) $>$ 0,4. Interestingly, the authors showed a direct correlation between the algorithm diagnostic performance and the number of images used for training. This outcome is not surprising, and clearly highlights the importance of training data, both in quantitative and qualitative terms, during the training phase of an algorithm. In fact, to date, the low number and small size of the available medically oriented datasets are the real bottlenecks that limit the development of clinically relevant computer vision algorithms. A similar approach was described by Xiong et al. [110], who developed a CNN-based diagnostic system using videoendoscopic images of laryngeal cancer, premalignant lesions, benign lesions, and normal tissue. The CNN detected lesions with an Acc of 87%, a Sen of 73%, a Spec of 92%, and an AUC of 92%. Moreover, the results were comparable to those obtained by a human expert with 20 years of experience.

Concerning real-time detection, Matava et al. [111] and Azam et al. [112] developed CNN algorithms applied in real-time during videoendoscopy and which aimed at identifying, on the one hand, normal airway anatomy and, on the other hand, UADT lesions. Using this type of approach, DL may be a useful complementary tool for clinicians in endoscopic examinations, progressively implementing the concept of human–computer collaboration. In detail, Matava et al. [111] compared the predictive performance of three models (ResNet, Inception, and MobileNet) in the identification of normal components of laryngeal and tracheal airway anatomy. ResNet and Inception achieved a Spec of 0.98 and 0.97 and a Sen of 0.89 and 0.86, respectively. Finally, Azam et al. [112] identified a CNN model for real-time laryngeal cancer detection in white light and NBI videoendoscopies. The dataset, consisting of 219 patients, was tested with an algorithm that achieved 0.66 Precision (Prec), 0.62 Rec, and 0.63 Mean Average Precision (mAP) with an IoU$>$0.5. In addition, the model ran with an average computation time per video frame of 0.026 s.

Segmentation Automated segmentation of anatomical structures in medical image analysis is a prerequisite for autonomous diagnosis and represents one of the most complex tasks in the field of computer vision. In this case, the algorithm not only needs to detect lesions but also needs to automatically delineate their margins as in Fig. 2.4. Recent CNN-based methods have demonstrated remarkable results and are well-suited for such a complex task.

During transoral laser microsurgery, a seven-class (void, VF, other tissue, glottic space, pathology, surgical tools, and tracheal tube) dataset was trained by Laves et al. [113] using a CNN-based algorithm. Different CNN architectures were investigated, and a weighted average ensemble network of U-Net and ErfNet (two of the most commonly used CNNs) turned out to be the best suited for laryngeal segmentation, with a mean IoU of 84.7%. Advances in ML and computer vision have led to the development of methods for accurate and efficient real-time segmentation. Paderno et al. [114] explored the use of fully CNNs for real-time segmentation of squamous cell cancer in videoendoscopies of the oral cavity and oropharynx. In this work, the authors compared different architectures and detailed their diagnostic performance and inference time, demonstrating their significant potential and the possibility of achieving real-time segmentation. However, for the first time, they suggested that highly heterogeneous subsites such as those encountered in the oral cavity may have inferior results when compared with more structurally homogeneous areas such as the oropharynx. This is in line with what was previously observed when applying bioendoscopic tools alone in a non-AI environment by Piazza et al. [106] and is possibly related to the larger epithelial differentiation within the oral cavity vs. the oropharynx and to specific limits related to oral examination (the presence of light artifacts and confounders such as tongue blade, teeth, or dentures).

When dealing with laryngeal lesions, Fehling et al. [115] explored the possibility of achieving a fully automated segmentation of the glottic area and VF tissue using a CNN in high-speed laryngeal videos. The algorithm obtained a Dice Similarity Coefficient (DSC) of 0.85 for the glottis, 0.91 for the right, and 0.90 for the left VF. Furthermore, the results revealed that, in both pathologic and healthy subjects, the automatic segmentation accuracy obtained was comparable or even superior to manual segmentation.

Generally, laryngo-pharyngeal lesions are those more frequently examined when measuring the role of automatic analysis by ML. In fact, only limited studies on nasopharyngeal disease differentiation have been performed on the basis of endoscopic images. For example, Li et al. [116] proposed a method to segment nasopharyngeal malignancies in endoscopic images based on DL. The final model reached an accuracy of 88.0%.

Finally, DL proved to be a promising addition to the field of endoscopic laryngeal high-speed videos. In clinical practice, the previous lack of dedicated software to analyze the data obtained resulted in a purely subjective assessment of the symmetry of VF movement and oscillation. The development of easy-to-use DL-based systems that are capable of automatic glottal detection and midline segmentation allowed obtaining objective functional data without the need for manual or semiautomatic annotation as previously described, among others, by Piazza et al. [117], thus significantly simplifying the process. These results were obtained through an organized and stepwise approach headed by the Erlangen research group that achieved high-fidelity automatic segmentation of the glottis [115] and glottal midline [118] as well as extraction of relevant functional parameters [119]. Thanks to these preliminary data, a DL-enhanced software tool for laryngeal dynamics analysis was developed [120]. This software provides 79 unique quantitative analysis parameters for video- and audio-based signals, and most of these have already been shown to reflect voice disorders, highlighting its clinical importance.

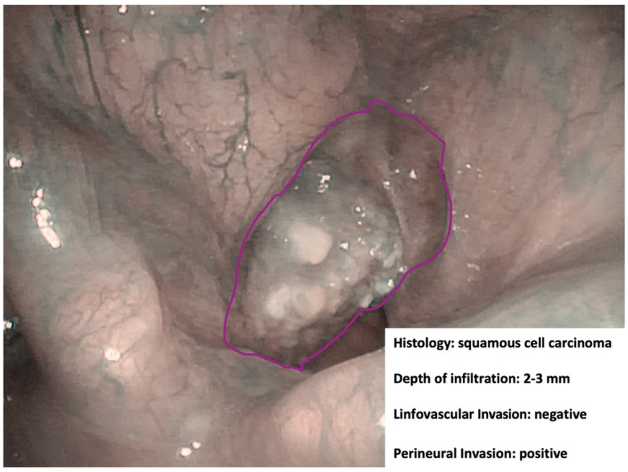

Histology: squamous cell carcinoma
Depth of infiltration: 2-3 mm
Linfovascular Invasion: negative
Perineural Invasion: positive

Figure 2.5: Endoscopic Narrow Band Imaging (NBI) frame showing an example of adjunctive data drawn from in-depth characterization by hypothetical machine learning algorithms.

In-depth characterization All the previously described tasks aim to provide an accurate definition of a given lesion, classifying it according to its nature, defining its location in the frame, and delineating its margins (with possible future roles in real-time definition of resection margins during a surgical procedure). However, all these objectives reproduce only what is generally achieved by an expert clinician and do not try to overcome the limits of human perception, even though their future implementation within a telemedicine environment would represent a large step toward more homogeneous diagnostic opportunities.

However, there is already indirect evidence that pattern recognition capabilities of novel AI systems may allow finding a correlation between the endoscopic appearance of a given lesion and its finer characteristics as represented in Fig. 2.5. Among these, depth of infiltration, so far investigable only by radiologic imaging or histopathologic evaluation [121], plays a remarkable role in the prognostication of oral cavity cancer and has fueled great interest in the possibility of speeding up its definition by AI tools applied to videomics. Identification through videomics of other tumor characteristics, such as histopathological risk factors (e.g., perineural and lymphovascular invasion), viral status (human papilloma and Epstein–Barr viruses), and genomic markers, is definitively more ambitious but already within the reach of similar approaches like radiomics and pathomics. Bridges connecting all these sources of information would be of great help in the near future to build up sharable profiling of tumors and their microenvironment.

Recent studies in the gastrointestinal tract, for example, have provided the proof of concept of this hypothesis and demonstrated that CNNs can differentiate between early and deeply infiltrating gastric cancer [122]. Nakahira et al. [123] further confirmed the potential of this approach by showing that CNN was able to correctly stratify the risk of gastric tumor development by analyzing the non-neoplastic mucosa at videoendoscopy.

Future Persectives The introduction of computer vision in UADT endoscopy is still in its infancy and further steps will need to be taken before reaching widespread application. In this view, the first step outside of purely research-driven applications will be the use of ML algorithms for human–computer collaboration. Dedicated algorithms can assist in every step of the endoscopic diagnostic approach, from quality assurance, effective storage and video classification, to risk determination, histologic definition, margins evaluation, and in-depth lesion profiling. As previously stated, this will be a stepwise approach that will start from easier tasks (i.e., quality assurance) and will progress toward more complex and more clinically relevant objectives. The ideal outcome will be to achieve accurate lesion characterization in terms of histologic nature, margins, and biologic characteristics and to be able to fully and objectively integrate these insights with data from other types of examinations (e.g., radiology, molecular biology, and histopathology).

Morphologic image analysis is the main field in which videomics is evolving in the context of clinical endoscopy. However, other more innovative aspects can be assessed by taking advantage of current computer vision technologies. A particularly interesting feature in otolaryngology is VF motility; in fact, an objective evaluation of this variable can be extremely helpful in both the assessment of functional deficits and in the precise staging of neoplastic disease of the glottis. This is especially true when considering that the international standard classification [124] of laryngeal cancer relies on purely subjective definitions of “normal vocal cord mobility”, “impaired vocal cord mobility”, and “vocal cord fixation” for the categorization of T1, T2, and T3 glottic tumors, respectively.

In this field, Adamian et al. [125] recently developed an open-source computer vision tool for automated VF tracking from videoendoscopies that is capable of estimating the anterior angle between VF of subjects with normal mobility and those with unilateral VF paralysis. The authors demonstrated the possibility of identifying patients with VF palsy by assessing the angle of maximal glottic opening (49° vs. 69°; $p < 0.001$). In particular, an angle of maximum opening $<58.6°$ was predictive of paralysis with a Sen and Spec of 0.85. Notwithstanding, this approach places significant limits on the evaluation of reduced mobility due to neoplastic involvement since it relies on the identification of the free margin of VF, which is often altered by glottic tumors. However, the development of alternative strategies is providing valuable outcomes in such a task.

Finally, novel surgical technologies such as transoral robotic [126] and exoscopic surgery [127] rely on digital video acquisition of a large amount of data and will potentially extend the applications of videomics to the intraoperative setting of quality and safety control as well as didactic proficiency. This is especially interesting considering the urgent need for more extensive training and collaborative datasets that will enable better refinement of ML algorithms, coming not only from diagnostic instrumentation but also from surgical robots and exoscopic tools.

2.2 A deep learning approach for keypoints localization for vocal folds movement assessment in endoscopic images

This section describes an innovative method for the automatic assessment of VF motility, based on the detection of five keypoints located at specific sites of the larynx in endoscopic images. VF motility assessment is not trivial, and while endoscopic imaging represents the gold standard for VF movement assessment, relying on visual examination of videoendoscopy is subjected to intrinsic challenges inherent to this imaging technique including the subjectivity of observers and the absence of standardized protocols, other than a range of technical and practical issues such as anatomical and image quality variations, motion artifacts and limited field of view.

Thus, using archived laryngoscopic images obtained from 124 patients, this section tackles the challenge of automatically detecting five keypoints and estimating VF motility through the development of a fully automated DL method. The architecture is designed as a heatmap regression network to enable precise estimation of the landmarks, thus facilitating the motility estimation. This study showcases the capability of tackling the challenge of keypoints detection in endoscopic images using DL, setting a foundation for subsequent exploration in this field. A further step for the estimation of VF motility will be delineated in Section 2.3. In terms of clinical utility, the proposed framework emerged as a valuable addition, improving the uniformity of VF motility estimation, increasing reproducibility, and reducing subjectivity and time needed for the assessment.

2.2.1 Introduction

With the advent of AI, the last decade has seen a revolution in the field of medical image analysis, with applications ranging from diagnosis to treatment, guidance, and follow up [53]. While promising results were obtained for processing anatomical images, such as CT [128], US [67], and MRI [129], as evidenced in Sec. 2.1, the analysis of endoscopic videos still represents a challenge [94] and only few commercially-available solutions exist [130]. This may be explained considering the peculiar challenges of endoscopic videos, including poor contrast, low signal-to-noise ratio, presence of motion blurring, and tissue motion. The field of otolaryngology and head and neck surgery makes not an exception [131]. Videoendoscopy is largely used in clinical practice for a number of applications, among which the assessment of VF motility.

VF are muscular structures located in the larynx, which are responsible for vocalization, breathing, and airway protection. Neurological and inflammatory diseases can lead to impaired VF movements [132], and the consequent paralysis of one or both VF may jeopardize key physiological functions of the larynx [133].

Diagnostic and therapeutic assessment of VF paralysis and paresis, glottic stenosis, and other neurologic disorders of the larynx are guided by videoendoscopic imaging [134].

Reliably monitoring therapeutic outcomes is crucial for maintaining high standards in clinical practice. VF movement disorders are identified by irregularities in the adduction and abduction of the folds. However, the prevalent outcomes in this area are often based on indirect measures of VF movements, using tools that evaluate voice quality and swallowing capabilities [135]. Currently, clinicians ascertain VF paralysis through folds position, using basic ordinal scales (like median, paramedian, or lateralized) in a static manner during vocalization. The clinical diagnosis of VF motility deterioration relies on the subjective examination and interpretation of VF motion during real-time viewing or playback of videoendoscopies. This evaluation is time-consuming and requires a skilled professional to be performed, and it is characterized by high inter- and intra-rater variability [136]. In this context, DL has the potential to tackle the variability of videoendoscopic frames and to provide clinicians with a quantitative assessment of VF motility.

Several studies measure the Anterior Glottic Angle (AGA) as a metric for understanding VF movements during VF abduction and adduction via laryngoscopy video. Traditionally, the AGA has been assessed by manually marking the laryngoscopy videoframes during standard inhalation and vocalization [137], during cough [138], and during specific lung function tests [139]; or using classic image processing approaches [140, 141]. While these traditional methods of image processing have been the standard for assessing the AGA, recent advancements in DL are paving the way for more sophisticated and automated approaches. Thus, VF motility has been assessed in terms of glottal segmentation, gauging motility through fold movement relative to the midline [142, 143]; or

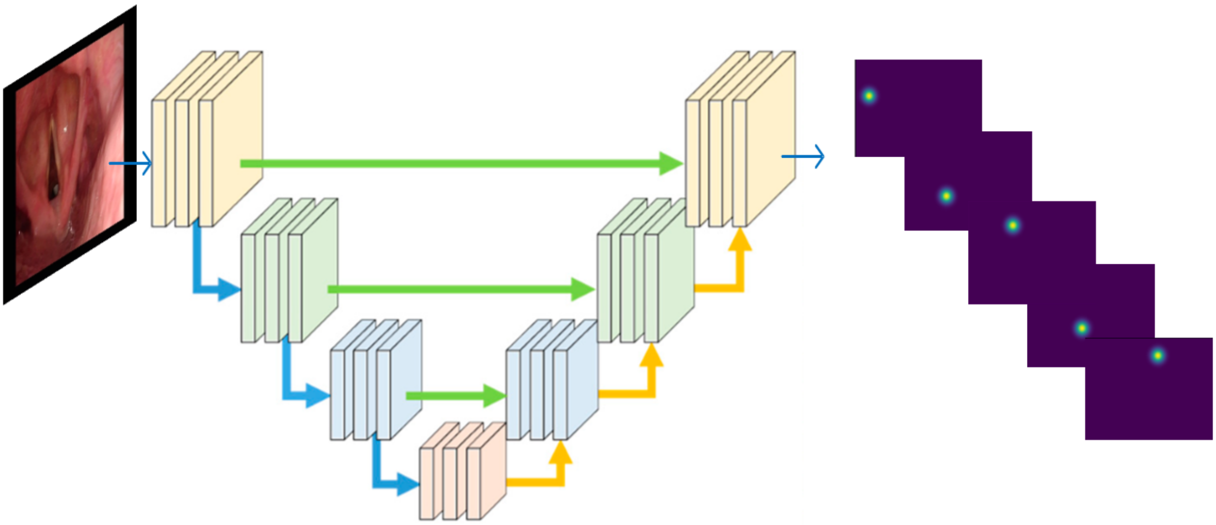


Figure 2.6: Representation of the proposed keypoints heatmap regression network.

through region of interest detection and glottal gap delimitation [144]. Glottal area segmentation for frame-wise estimation of the AGA was also used in [145], for VF tracking from laryngoscopy video.

More recently, VF motility has been assessed in terms of anatomical keypoints (or landmarks) tracking. Keypoints estimation from endoscopic laryngeal images may be crucial to provide quantitative measurements to ensure objective analysis of VF motility, supporting diagnosis and treatment planning. The use of keypoints was first explored in [146, 147] to estimate the AGA. In [146] an open-source DL toolbox (DeepLabCut) was employed to train a computer vision model for offline localization of the free edges of VF from laryngoscopy video examinations.

To estimate VF keypoints, two common approaches can be the direct coordinates regression and the heatmaps regression: while direct coordinate regression aims to predict the exact (x, y) coordinates of the keypoints, heatmaps regression offers a different perspective, allowing for the detection of keypoints presence and their general location. Considering that heatmaps may provide better robustness to partial occlusion, a valuable feature in videoendoscopy where VF and other anatomical landmarks may be partially obscured, this work proposes the first DL algorithm for VF motility assessment through keypoints detection based on heatmap regression.

2.2.2 Materials and methods

This section introduces the proposed framework, the datasets used, and the training settings. Fig. 2.6 shows an overview of the proposed keypoint detection network for VF motility estimation.

Dataset description

The dataset used in this study is made of videoendoscopic frames of patients treated at the Unit of Otorhinolaryngology-Head and Neck Surgery, University of Brescia, Italy. Data were acquired following the principles of the Helsinki Declaration, and approval was obtained by the local ethical committee. A total of 471 endoscopic images from 124 patients (28 of which oncologic) were collected from a dedicated archive and anonymized. For each video, a variable number of representative frames were selected and annotated using Label-studio by an expert laryngologist with more than ten years of experience. Annotation consists of 5 keypoints located at specific sites of the larynx, fundamental for VF motility estimation: the epiglottic insertion point of the left and right aryepiglottic folds (Left Epiglottic (LE), Right Epiglottic (RE)), the posterior angle of the left and right vocal folds (Left vocal fold (LV), Right vocal fold (RV)), and the Anterior commisure (A), as shown in Fig. 2.7. The choice of these keypoints is driven by clinical considerations: RE and LE mark the aryepiglottic fold's insertion into the epiglottis, serving as pivotal points fixed during arytenoid movement, making them apt references for appraising supraglottic larynx motion; LV and RV can help assessing the degree of VF closure during phonation; A is a pivotal point for understanding VF dynamics, as changes in its position and movement can reveal information about VF tension and adjustments made during voice production.

To cope with the small amount of data, and to effectively use all the data available, a five-fold cross-validation was performed. The dataset was split into five balanced subsets, in each fold four subsets were used for training and validation, while the remaining for testing. For each fold, images were selected to ensure no patient overlap between the train and test sets.

Proposed method

The proposed model is inspired by the classical encoder-decoder architecture of U-Net [72]. The MobileNetV2 architecture pre-trained on ImageNet was employed as the encoder $e(\cdot)$, which serves as a feature extractor. Additionally, a decoder network $d(\cdot)$ was used to recover spatial information and generate the heatmaps. The $e(\cdot)$ is composed of an initial convolutional layer with 32 filters, which reduces the image size by half, followed by batch normalization and a Rectified Linear Unit (ReLU) activation function. This initial layer is followed by a series of inverted residual blocks consisting of an initial 1x1 convolution followed by a 3x3 depthwise convolution and ending with another 1x1 convolution. At each block, the number of channels increases enabling the incremental learning of more complex features. The number of channels starts from 32 and progressively increases to 576. Similarly, the $d(\cdot)$ is composed of four blocks, each comprising two 2D conv layers followed by a ReLU activation function and batch normalization. To recover the lost features resulting from downsampling in the $e(\cdot)$ path, the input of each block is concatenated with the corresponding feature maps from $e(\cdot)$. The last block consists of three 2D conv layers, with the first two being followed by a ReLU activation function, and the last one activated by Softmax.

The proposed CNN is fed by stacking the endoscopic frames and the five corresponding heatmaps of dimension W x H, where W and H represent the width and height, of the endoscopic images,

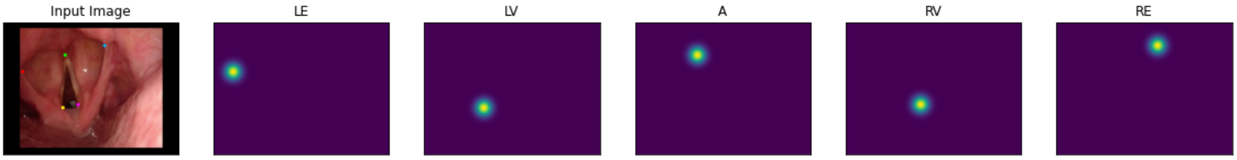


Figure 2.7: Visual sample of a labeled image. The colored points in the leftmost image represent the keypoints: left epiglottic (LE) in red, left vocal fold (LV) in yellow, anterior commissure (A) in green, right vocal fold (RV) in magenta, right epiglottic (RE) in cyan. The generated heatmaps for each keypoint are also reported separately.

respectively. Each heatmap is represented by a Gaussian distribution with a standard deviation (σ) equal to 20 and centered at the keypoints center.

Training settings

All frames were resized to 224x224 pixels, and mean intensity was removed from each frame. The model was trained for 200 epochs, optimized using the Adam optimizer with an initial learning rate of 0.001, with a batch size of 8. In the proposed approach, a Channel-weighted Mean Square Error (CW-MSE) loss function was introduced to give differential importance to individual channels of the target tensor. The CW-MSE loss function, L, is formulated as follows:

$$L = \sum_{i=1}^{N} \left(\text{mean} \left(\text{w}[i] \times (y_{\text{pred}}[..., i] - y_{\text{true}}[..., i])^2 \right) \right)$$

Where:

- N is the number of channels.
- $y_{\text{pred}}[..., i]$ and $y_{\text{true}}[..., i]$ denote the i^{th} channel of the predicted and true tensors, respectively.
- w is a list of scalar values, where each value corresponds to the weight of a specific channel.
- The mean operation computes the average of the squared weighted differences for each channel.

The intuition behind the CW-MSE loss is to allow the model to focus more on channels that are deemed more critical for the task at hand. By assigning higher weights to these channels, the model can be guided to produce more accurate heatmaps for them.

During the training, on-the-fly data augmentation was performed to enhance generalization performance. The augmentation techniques included geometrical transformations such as horizontal and vertical flipping, and random rotation in the range of ± 30 degrees, and intensity transformations such as random brightness correction, random hue adjustment, and random saturation. These augmentations were randomly applied at each training iteration. The best model among epochs is selected based on the lowest loss value obtained on the validation set.

All the analysis were performed using Tensorflow 2.x on an NVIDIA RTX 2080 TI, with a Xeon e5 CPU and 128 GB RAM.

Comparison with literature

Approaching keypoints estimation through a heatmap regression network, instead of a direct coordinates regression network, was driven by previous work from different fields [148, 149], which showed that deducing joint positions from an input frame (and thus from direct coordinates regression) is notably non-linear. The proposed regression system, instead, generates stacked confidence maps each having the same size of the input frames (i.e. W x H).

Nevertheless, a comparative analysis of the proposed model with a direct regression model was conducted to prove the effectiveness of the development of a heatmap regression network rather than a direct keypoint coordinates regression approach.

The model used for direct coordinates regression is a fully CNN, made of a MobileNetV2 pre-trained on ImageNet as backbone to extract meaningful features from the images which are later

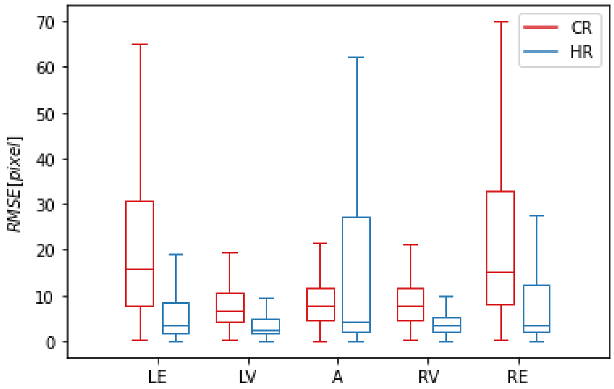


Figure 2.8: Boxplots of the Root Mean Square Error (RMSE) for each of the five keypoints: left epiglottic (LE), left vocal fold (LV), anterior commissure (A), right vocal fold (RV), right epiglottic (RE). The proposed heatmap regression model (HR), is compared with the direct coordinate regression model (CR). Each boxplot represents the distribution of the RMSE values across all the folds.

passed to a custom regression head made of two separable convolutions for predicting the keypoints coordinates. The backbone used in this model is the same architecture used in the encoder path of the proposed heatmap regression method.

For a fair comparison, the two models are trained under the same settings in five-fold cross validation and with the same computational resources.

The keypoints detection performance of the two models was evaluated based on the Root Mean Square Error (RMSE) defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - G_i)^2} \tag{2.1}$$

where N is the number of keypoints, P_i represents the coordinates of the predicted location of the $i-th$ keypoint, G_i represents the coordinates of the ground truth location of the $i-th$ keypoint. The difference (P_i - G_i) is computed as the Euclidean distance between the predicted and ground truth positions for each keypoint. RMSE values are expressed in pixels.

For the heatmap regression model, the RMSE was evaluated considering the maximum value of the predicted and ground truth heatmaps.

2.2.3 Results

The proposed method obtained mean RMSE values over the 5 folds equal to 16.44, 4.93, 20.12, 5.70, 21.93 pixels for LE, LV, A, RV, and RE respectively. The direct coordinate regression method, instead, achieved mean RMSE values over the 5 folds of 22.70, 8.50, 9.39, 9.35, 22.91 pixels for LE, LV, A, RV, and RE respectively. Average results for each fold of the proposed keypoint detection methods are reported in Table 2.1, in comparison with the outcomes of the direct coordinate regression method.

In Fig. 2.8 the boxplots relative to the RMSE for each keypoints are reported for the two models, while visual samples of the qualitative results can be seen in Fig. 2.9, where it is possible to appreciate the improvements brought by the proposed heatmap regression model, which are particularly evident in the accurate positioning of keypoints with respect to the other model.

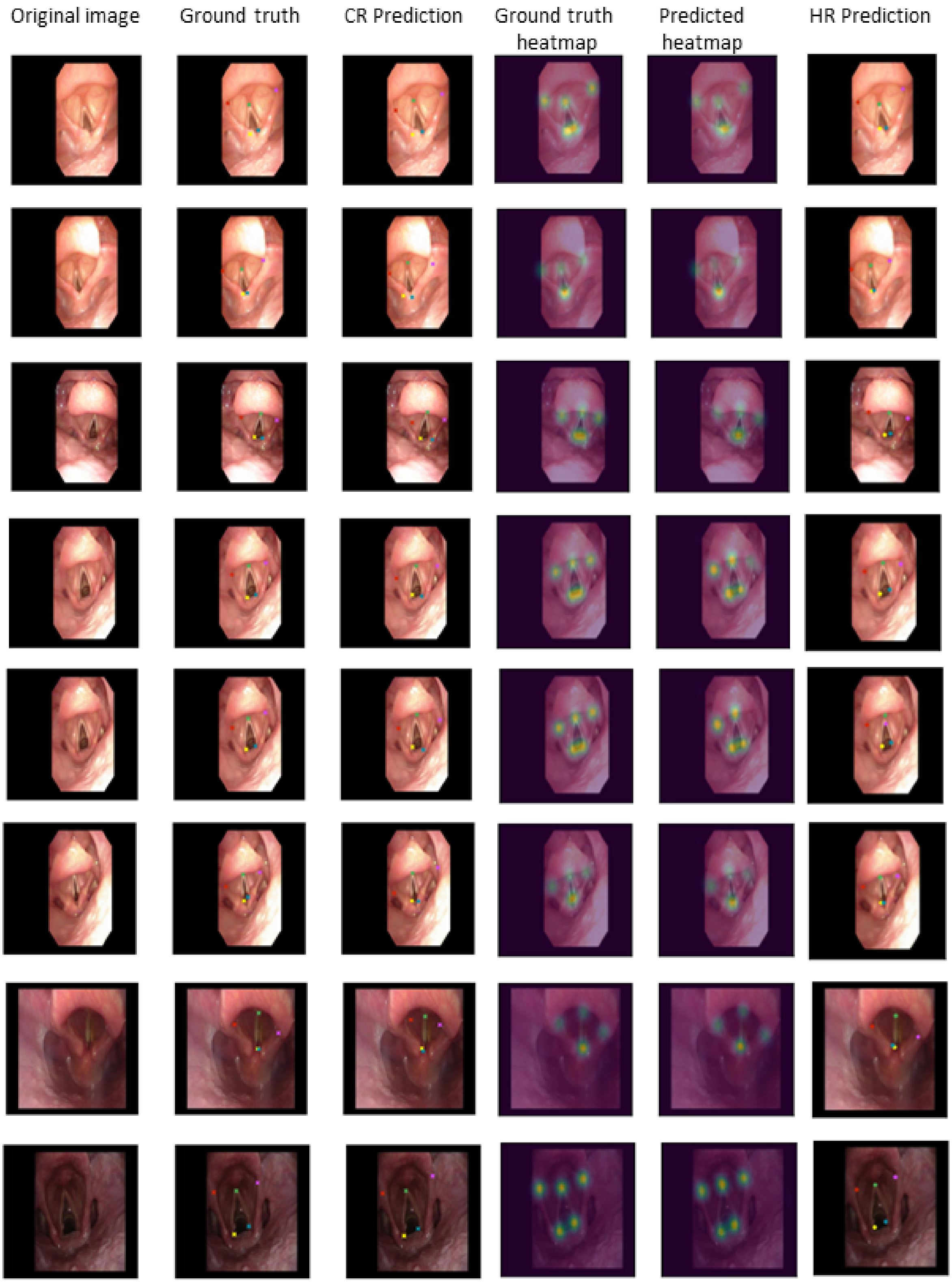


Figure 2.9: Qualitative results from the comparison of the two models on sample test images. From left to right: original image, ground truth coordinate annotation, prediction of the direct coordinate regression model (CR), ground truth heatmap, predicted heatmap, prediction of the heatmap regression model (HR).

Table 2.1: Results of the root mean squared error (RMSE) computed on the test sets of the five-folds. Values are expressed in pixels.

Fold	Model	LE	LV	A	RV	RE
F 1	Coordinates regression model	20.79	8.40	8.85	8.07	20.60
	Heatmap regression model	13.71	4.39	63.67	4.13	14.93
F 2	Coordinates regression model	26.05	9.04	8.84	10.10	25.15
	Heatmap regression model	25.27	5.73	7.90	7.08	23.04
F 3	Coordinates regression model	21.66	8.99	9.06	9.06	20.92
	Heatmap regression model	7.72	5.47	7.38	5.42	38.23
F 4	Coordinates regression model	25.26	8.15	9.71	9.18	25.90
	Heatmap regression model	14.25	4.48	13.36	5.17	21.24
F 5	Coordinates regression model	19.74	7.91	10.50	10.35	21.99
	Heatmap regression model	20.98	4.59	8.30	6.70	12.25
Average	Coordinates regression model	22.70	8.50	9.39	9.35	22.91
	Heatmap regression model	16.44	4.93	20.12	5.70	21.93

2.2.4 Discussion

Despite the growing interest in assessing VF motility and the established value of endoscopic analysis, there are still issues to be faced. These challenges stem from the high expertise required to perform and interpret the procedure, the lack of standardization, and the high variability among endoscopists' evaluations. To address these issues, a DL approach is proposed to detect five keypoints located at specific locations of the VF, which is a preliminary but fundamental task to estimate VF motility.

From Fig. 2.8 a certain difference can be noticed among the five keypoints. Particularly, when compared to the others, the two external keypoints (LE and RE) exhibit the highest error on average for the direct coordinate regression approach. This behaviour is possibly related to the higher variability in the position of the external-most keypoints, and to the occlusion caused by other anatomical structures. The prediction of these two keypoints is highly improved using the heatmap regression approach, thanks also to the use of the CW-MSE loss which improves the prediction of these two external-most keypoints. Only for the regression of the keypoint A, the proposed model achieves lower performance with respect to the direct coordinate regression model. Results obtained from the comparison of the two models show that direct coordinates regression results in a higher median RMSE for all the keypoints, confirming the advantage of regressing heatmaps for capturing nuanced spatial relationships and potentially yielding more robust predictions.

Even if the achieved results are promising, a limitation of this work can be seen in the fact that the analysis only included a limited number of videondoscopies of patients for whom the VF motility was preserved. Nevertheless, the study can be extended also to patients that are in conditions of reduced motility.

2.2.5 Conclusion

This section outlined a keypoints detection model for VF motility estimation in videoendoscopic images based on heatmap regression. The results achieved on a newly collected dataset suggest that keypoints detection based on heatmap can be successfully exploited to estimate VF motility, obtaining better performance with respect to direct coordinate regression. Thus, the proposed solution moves the state of the art towards a better framework for VF motility assessment and can lead to applications in computer-aided diagnosis.

2.3 Classifying vocal folds fixation from endoscopic videos with machine learning

The method reported in this section is a natural progression from the work described in the previous Sec. 2.2; indeed, it demonstrates that VF motility estimation can be derived from the keypoints coordinate predictions obtained with the previous method. Here, however, the estimation is conceptualized based on the ground truth coordinates.

A conference paper on this work has been presented at EMBC 2023 and published as [150]: Villani, F.P., Paderno, A., Fiorentino, M.C., Casella, A., Piazza, C., Moccia, S. (2023). Classifying Vocal Folds Fixation from Endoscopic Videos with Machine Learning. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2023:1–4.

2.3.1 Introduction

As stated before, the clinical diagnosis of VF motility relies on the subjective examination and interpretation of VF motion during real-time viewing or playback of videos captured through videoendoscopy. This evaluation is time-consuming and requires a skilled professional to be performed, and is characterized by high inter- and intra-rater variability [136]. In this context, ML has the potential to tackle the variability of videoendoscopic frames and to provide a quantitative perspective to the analysis of VF motility. Thus, this section focuses on the analysis of endoscopic frames extracted from endoscopic videos, proposing a ML algorithm for the assessment of VF motility.

The literature on ML algorithms for laryngeal videoendoscopic image analysis has been growing since 2017. The work in [151] is among the first to investigate the use of ML algorithms for early stage cancerous laryngeal tissue classification. Since then, several studies have been published, including a recent review [131]. Motility assessment, instead, is mostly addressed with DL methods based on glottal segmentation, from which the motility is evaluated based on the movement of each fold with respect to the midline; or through region of interest detection, and glottal gap delimitation [144]. Hamad et al. [142] developed a DL system for automatic segmentation of the glottal region in laryngoscopy videos using a fully convolutional regression network. More recently, Yousef et al [152] studied VF kinematics during the running speech, analyzing VF vibrations in adductor spasmodic dysphonia. A U-Net was deployed for glottal area segmentation in high-speed videoendoscopy to quantitatively analyze vibrations in both healthy and unhealthy patients. Similarly, in [143] VF dynamics is evaluated in association with voice disorders. They trained a deep neural network with data from laryngeal high-speed videoendoscopy with the aim of segmenting the glottal area, from which the glottal edges are derived during connected speech. Other studies make use of phasegram [153] (a visualization method of system dynamics that can be interpreted as a bifurcation diagram in time) or phonovibrogram [136, 154] (a graphical representation of the VF deflections, automatically extracted from laryngeal high speed recordings) to evaluate VF motility related with voice disorders. However, unlike videoendoscopy, these kinds of tests are not usually

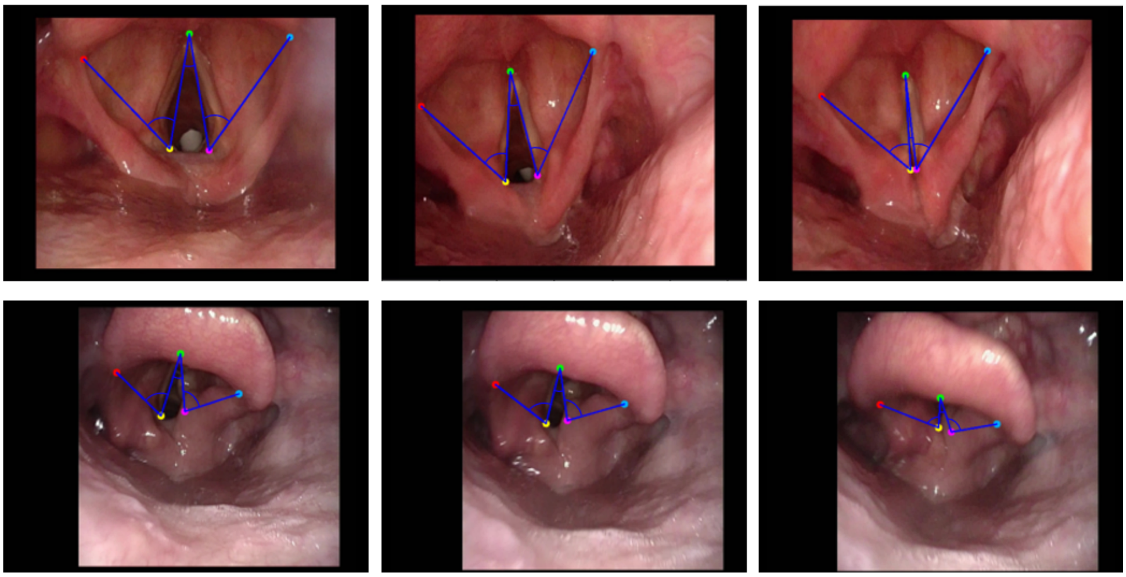


Figure 2.10: Representation of three consecutive frames (from left: abducted, normal, and adducted vocal cords, respectively) with ground truth keypoints annotation. The images in the first row refer to a subject with preserved motility, while the ones in the second row to a subject with fixation. The colored points represent the keypoints: left epiglottic in red, left vocal fold in yellow, anterior commissure in green, right vocal fold in magenta, right epiglottic in cyan.

performed in clinical practice.

Differently from the work in the literature, this work relies on ML for VF motility estimation, proposing a method to classify motility into two classes (namely: preserved motility and fixation) based on keypoints. This method is advantageous as it allows to directly obtain a classification, without the need of post-processing, as in the case of glottal segmentation. Each of the selected keypoints represents an important clinical landmark for the analysis, providing a close approximation of both glottic and arytenoid movements. Starting from the coordinates of the five keypoints, clinically relevant features were handcrafted to train the classification models.

2.3.2 Materials and methods

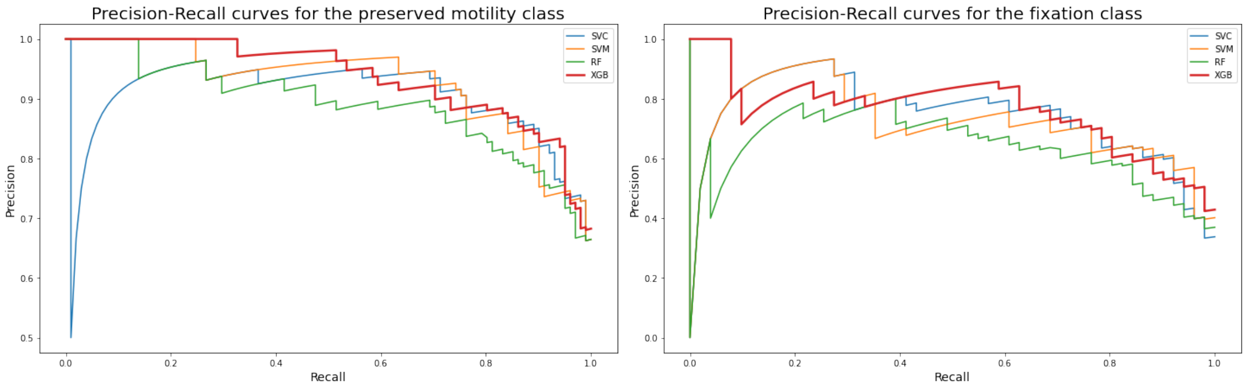
Vocal Folds Model and Keypoints Annotation

The dataset used for this analysis is made of videoendoscopic frames of patients treated at the Unit of Otorhinolaryngology - Head and Neck Surgery, University of Brescia, Italy. Data were acquired following the principles of the Helsinki Declaration, and approval was obtained by the local ethical committee of Spedali Civili of Brescia. A total of 558 endoscopic images from 186 patients were collected from a dedicated archive and anonymized, and for each video three representative frames were selected. The motility was estimated among these three endoscopic frames from five keypoints chosen according to the clinical experience of the clinicians, and located at specific sites of the larynx, as previously detailed: LE, RE, A, LV, and RV, as shown in Fig. 2.10.

Frames annotation was performed by an expert (more than 10 years of experience) laryngologist using LabelMe[1]. Only subjects for which three frames representing a specific VF position (abducted, neutral, adducted) were available, were included in the study. After this process of data selection,

---

[1] https://github.com/wkentaro/labelme



(a) **(b)**

Figure 2.11: Precision-Recall curves calculated on the test set, for all the classifiers. The best-performing classifier resulted to be XGBoost, showing the highest average precision and area under the precision-recall curve for both classes.

the collected dataset counted 101 subjects with preserved motility and 51 subjects with fixation. The dataset includes both oncologic and non-oncologic patients.

Features Extraction and Classification

To assess VF motility, the following features were extracted from the labeled frames:

- The central and the two external angles for each frame (as shown in Fig. 2.10).
- The static index: the difference between the two external angles for each of the three frames.
- The dynamic index: the ratio between the difference of the right angle in the first and third frames and the difference of the left angle in the first and third frames.

Common ML classification algorithms were investigated and compared, including SVM with linear (SVC) and non-linear kernels, XGBoost (XGB), and Random Forest (RF). The optimal hyperparameters for each classifier were retrieved via grid-search and cross validation on the training set, using stratified three-fold cross validation. This ensures that every patient in the dataset appears at least once in the testing set. In particular, the three-fold cross validation cyclically splits the dataset into three equally sized folds, of which two are used to train and one to validate and tune the parameters. Before classification, features were normalized by removing the mean (centering) and scaling to unit variance. Given the unbalance between the two classes, the minority class was over-sampled using the synthetic minority oversampling technique (SMOTE). Also class weights were balanced according to the number of samples of each class.

Experimental Analysis

The performance of the classifiers was evaluated using classification Prec, Rec, and F1-score on the test set. Considering the unbalance of the dataset, the AUC and the Average Precision (AP) were also computed.

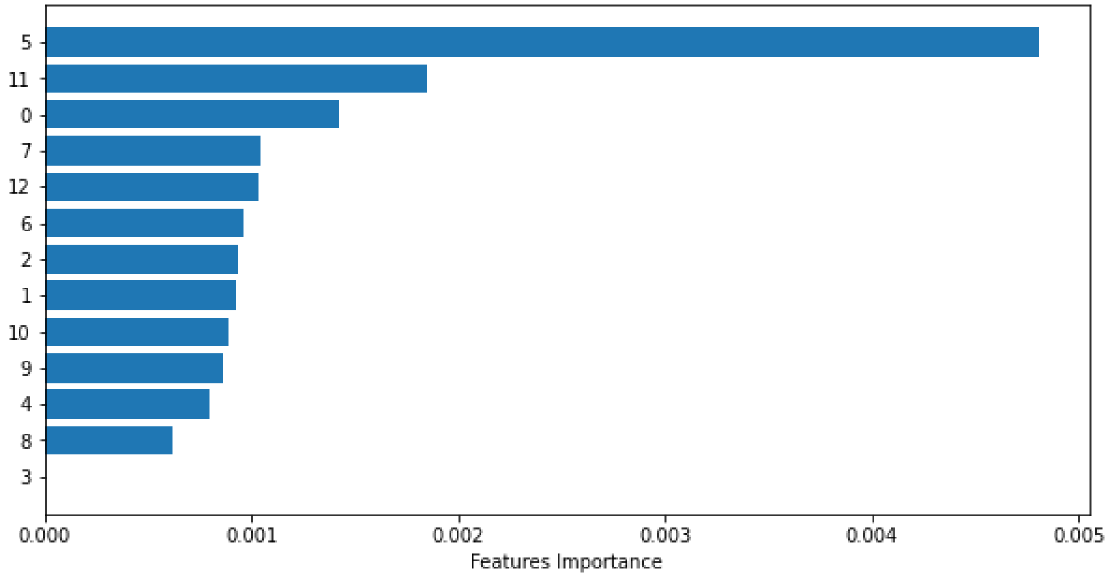


Figure 2.12: Features importance of the XGBoost classifier. Features from 0 to 8 refer to the three angles (central and externals) of the three successive frames, features from 9 to 11 refer to the static indexes of the three frames, and feature 12 refers to the dynamic index.

2.3.3 Results

The performance of all the classifiers is shown in Table 2.2, results are reported in terms of the metrics computed on the test set. Fig. 2.11 shows the Prec-Rec curves of all the classifiers. All the tested models showed comparable results, however, the best-performing classification algorithm resulted to be the XGB, with an AP of 0.76 and 0.94, and an AUC of 0.76 and 0.93 for the fixation and preserved motility class, respectively. Features importance of the XGB classifier is reported in Fig. 2.12. Specifically, on 152 test subjects (among the three cross validation folds), XGB achieved the lowest number of incorrect predictions (27 subjects). Samples of misclassified frames are shown in Fig. 2.13.

2.3.4 Discussion

The main objective of this study was to evaluate the ability of ML algorithms to discriminate between vocal cords preserved motility and fixation. To do so, a number of relevant features was extracted from triplets of videoendoscopic frames, representing specific VF positions. The extracted features were used to train and test four different classifiers, which showed good results, and the best-performing one resulted to be the XGB. Even though the results of this model do not depart from the others, the use of this specific ML classifier could be useful in the case of some not labeled keypoints, as it is able to handle missing values [155]. From the results, it is also possible to appreciate the ability of all the tested models to assess VF motility. This is an expected behavior [156, 157] and confirms that the application of ML may have a positive impact on assisting clinicians in their practice.

To the best of our knowledge, this is the first study to rely on keypoints to evaluate vocal cords motility. Previous work in literature, in fact, focused on the segmentation of the glottis to evaluate the motility. The advantage of relying on keypoints, as already demonstrated in precedent work from other fields [158, 159], is the possibility to obtain a direct classification. Methods relying on

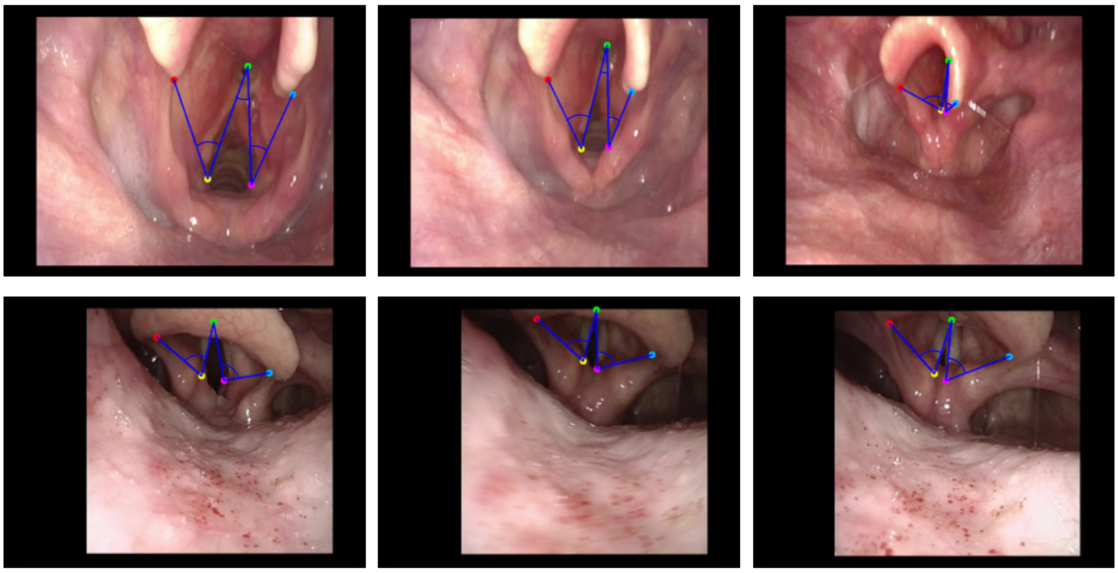


Figure 2.13: Visual samples of misclassified frames. The images in the first row were erroneously predicted as belonging to the fixation class, while the images in the second row were erroneously predicted as belonging to the preserved motility class. In the latter case, the vocal folds area occupies a small portion of the frame, which makes the prediction more challenging.

Table 2.2: Performance evaluation metrics. Precision (Prec), recall (Rec), F1-score (F1), accuracy (Acc), Average Precision (AP), and Area Under the Curve (AUC) are reported. For each classifier, the first row refers to the class fixation, while the second to the class preserved motility.

Classifier	Prec	Rec	F1	Acc	AP	AUC
	0.73	0.73	0.73		0.75	0.73
SVC	0.86	0.86	0.86	0.82	0.90	0.90
	0.71	0.76	0.74		0.72	0.71
SVM	0.88	0.84	0.86	0.82	0.93	0.92
	0.67	0.59	0.62		0.64	0.63
RF	0.80	0.85	0.83	0.76	0.89	0.89
	0.76	0.69	0.72		0.76	0.76
XGB	0.85	0.89	0.87	0.82	0.94	0.93

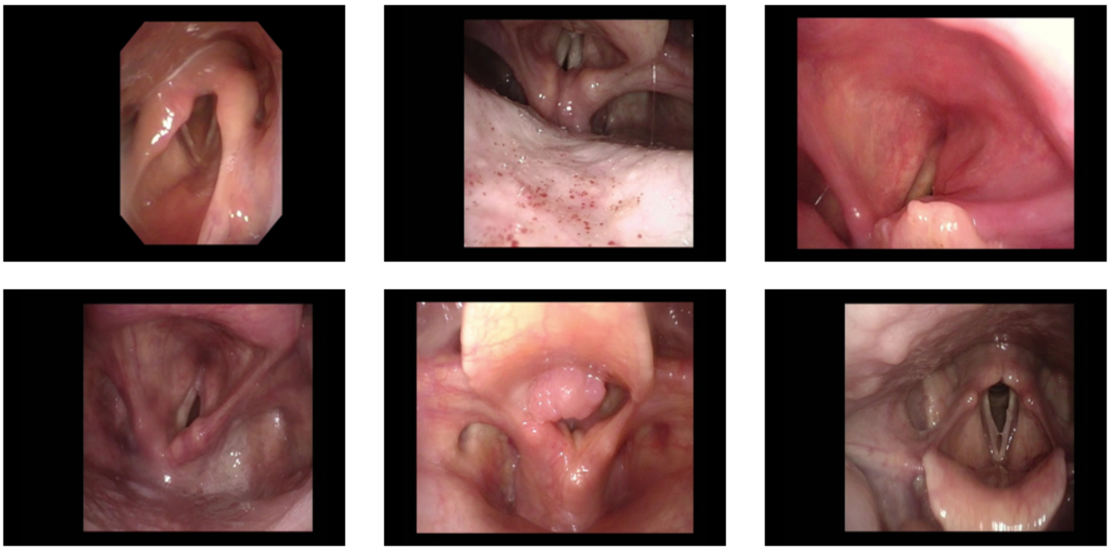


Figure 2.14: Visual samples of frames from the used dataset. It is characterized by high variability among the frames, which reflects also on the variability of the features used to train the models.

segmentation, in fact, need a post-processing step to obtain a diagnosis.

A limitation of the proposed work could be seen in the relatively limited size of the dataset, which is due to the time needed to label each frame, and to the lack of available annotated dataset online. The time consuming annotation procedure also makes it difficult, at the moment, to evaluate intra-observer variability. Moreover, the dataset used in this work includes frames with very high variability among each other, as shown in Fig. 2.14, which is typical of videoendoscopic frames. This characteristic of the dataset reflects also on the extracted features and on the achieved results. For this reason, adding the classification algorithm downstream of a frame selection process might improve the results.

As future work, to support clinicians in the actual clinical practice, the classification model could be included within other computer-assisted algorithms for diagnostic support, e.g., frames selection and automatic keypoints regression.

2.3.5 Conclusion

VF fixation is typically assessed by visually evaluating videoendoscopic frames. This process is time-consuming and requires an expert eye. To make the evaluation more objective, in this paper four ML models were compared to classify vocal cords motility into two classes: preserved motility and fixation. The best-performing model, XGB, proved to be a useful tool to investigate vocal cords motility in a more objective and reliable way. It is, in fact, able to distinguish between the two classes, which makes it a potential tool to support clinicians in their clinical practice.

2.4 Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes

In this last section, a different application within the realm of video endoscopy will be outlined. It pertains to the automatic cancer segmentation in the UADT from endoscopic images. As outlined in the previous sections, laryngoscopy represents a gold standard screening diagnostic tool for the diagnosis of precancerous lesions and early cancer of the larynx.

A journal paper on this work has been published as [160]: Paderno A., Villani F.P., Fior M., Berretti G., Gennarini F., Zigliani G., Ulaj E., Montenegro C., Sordi A., Sampieri C., Peretti G., Moccia S., Piazza C. (2023). Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. Acta Otorhinolaryngol Ital, 43(4), 283-290.

2.4.1 Introduction

The application of computer vision techniques in diagnostic videoendoscopies (i.e. videomics) [88, 161] is a promising research field that is currently showing a fast rate of growth in many medical specialties. The recent refinement of DL algorithms for image processing and their application in the medical field opened novel possibilities in the management of endoscopic exams that, in the past, had only subjective value. In particular, videoendoscopy is a key component in the management of UADT tumors, influencing their entire diagnostic process, treatment, and follow-up [131]. Notwithstanding, it remains an operator-dependent and time-consuming procedure, which is substantially limited by the variables of human experience and perception. This is especially true when endoscopy is applied in conjunction with optical biopsy techniques such as NBI [106], requiring even more specialized training and adding a further layer of complexity and subjectivity. Finally, no easily classifiable and structured data can be drawn from these examinations, significantly limiting their integration with other technologies (e.g., cross-sectional imaging, US, genomic markers, and so on). This is also highlighted by initial attempts to standardize endoscopic evaluation and improve the implementation of new analytic techniques [162]. This study aimed to explore the potential of a DL algorithm, Mask R-CNN [163], in the diagnostic approach to UADT Squamous Cell Carcinoma (SCC). The primary goal was to detect and classify neoplastic lesions and, at the same time, precisely define their margins, a task overall defined as instance segmentation. In fact, Mask R-CNN provides a flexible and general framework that can also be potentially applied to medical images. This approach combines elements from the tasks of object detection (where the goal is to localize the lesion using a bounding box), object classification [164] (where the purpose is to classify each pixel into a set of categories – e.g., tumor vs. normal mucosa), and semantic segmentation (where the aim is to automatically delineate the lesion's margins). Finally, three different areas of the UADT (oral cavity, oropharynx, larynx/hypopharynx) were included in the analysis in order to identify potential site-related differences in the diagnostic capability of the proposed DL algorithm, a piece of information that is still lacking in the current literature. In fact, studies assessing the value of AI in endoscopy are generally focused on a single site and are difficult to generalize in the context of UADT SCC, which can arise from a wide variety of anatomical structures, as well as epithelial and mucosal types.

2.4.2 Materials and methods

A retrospective study was performed including videoendoscopies performed between September 2009 and January 2021 in patients treated at the Unit of Otorhinolaryngology–Head and Neck Surgery, University of Brescia, Italy for SCC of the UADT. A total of 7567 videoendoscopies were collected from a dedicated archive. All recordings were anonymized and associated with the corresponding histopathology report.

The primary endpoint of this study was the definition of the diagnostic accuracy (in terms of DSC) of the Mask R-CNN model when applied to NBI UADT videoendoscopic frames. The secondary endpoint was the comparison of the algorithm's DSC in the three different anatomical areas herein considered. The inclusion criteria were as follows:

- Primary or recurrent SCC of the UADT (differentiated according to the anatomical site, namely: the oral cavity, oropharynx, and larynx/hypopharynx);
- NBI evaluation with adequate quality (without pooling of saliva, blood spots, swallowing reflex, coughing, or other technical issues);
- Available histological examination obtained at the time of videoendoscopy or subsequent surgery.

All patients were examined both under white light and NBI through transnasal videolaryngoscopy (HD Video Rhino-laryngoscope Olympus ENF-VH, ENF-VQ, or ENF-V2, Olympus Medical System Corporation, Tokyo, Japan) or through transoral endoscopy by a zero-degree rigid telescope coupled to an Evis Exera II HD camera connected to an Evis Exera II CLV-180B/III CV-190 light source (Olympus Medical Systems Corporation, Tokyo, Japan). Endoscopic videos were selected independently by two otolaryngologists with extensive experience (at least 4 years) in endoscopic assessment of UADT lesions by NBI and independently reviewed by an adjunctive expert. Images were then manually quality-controlled, with the exclusion of those that were blurred, obscured by blood or secretions, or without adequate NBI evaluation.

Image processing

Three representative frames per video were selected for every lesion: the most representative NBI videoframe was chosen and subsequent frames at 0.3 second time intervals were then automatically selected. Frame annotation was performed manually using LabelMe [165]. Annotation consists of a variable number of keypoints marking the lesion margins in the videoendoscopic frame taking into account positive NBI patterns. The resulting masks were then saved in JSON format and stored in a dedicated folder. Two clinical experts concomitantly annotated the images and a further review was performed by a senior staff member. When an agreement regarding lesion margins was not reached, the frame was excluded from the analysis. After this selection process, a total of 1034 endoscopic images were obtained. Three different sub-datasets were generated according to the lesion's primary site: oral cavity, oropharynx, and larynx/hypopharynx. In this way, the total frames analyzed were 653 for the larynx/hypopharynx, 246 for the oral cavity, and 135 for the oropharynx.

Dataset description

The dataset included 1034 images from 323 patients. For algorithm training and testing the dataset was split over patients and balancing the three classes into three sets: 935 images from 290 subjects for training, 48 images from 16 subjects for validation, and 51 images from 17 subjects for testing. All images were resized to the same dimension of 480 x 640 pixels.

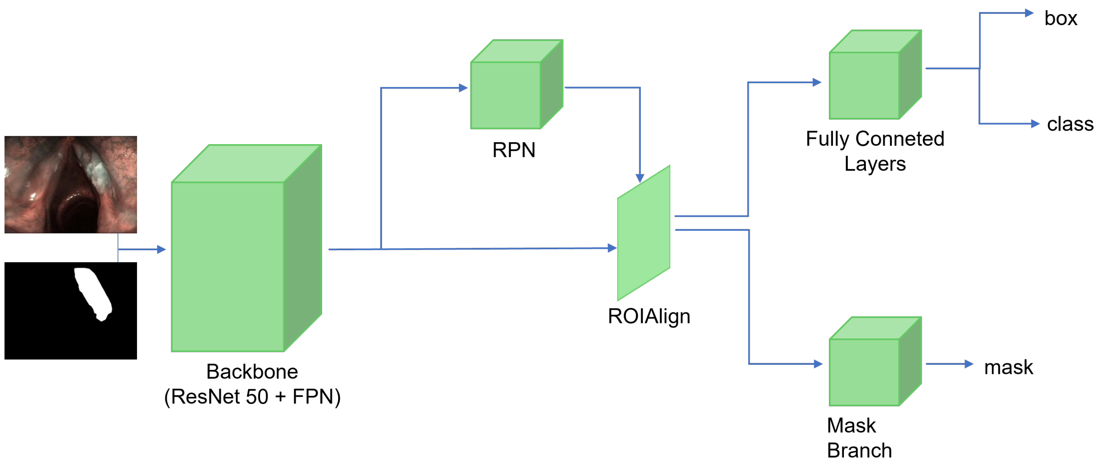


Figure 2.15: Schematic representation of the proposed architecture. The Mask R-CNN is made of a backbone (composed of a ResNet50 and a feature pyramid network), a Region Proposal Network (RPN), ROIAlign, and three heads, for classification, bounding-box regression, and segmentation.

Deep learning analysis

In this work, Mask R-CNN [166] was used to segment the tumor in endoscopic frames. This CNN consists of a backbone, Region Proposal Network (RPN), and three heads for classification, bounding-box regression, and segmentation (Fig. 2.15). As backbone, the ResNet50 [59] was used combined with a Feature Pyramid Network (FPN) [167] to extract features from the input frame at multiple scales. Starting from the features computed with the backbone, the RPN identifies candidate regions containing the tumor. For each of the proposed regions, the final bounding box containing the tumor and the tumor segmentation are obtained from the three heads. To cope with the relatively limited size of the dataset, the weights computed on the COCO dataset [168] were used to initialize the layers of the Mask R-CNN; and to reduce the risk of overfitting, on-the-fly data augmentation was performed during training by applying: random brightness changes in the range (0.5, 1.1), random contrast changes in the range (0.8, 3), and random rotation in the range (-20, 20 degree). The model was trained for 100 epochs, using the Stochastic Gradient Descent (SGD) as optimizer with an initial learning rate of 0.001 and momentum of 0.9. The loss function used to train the model is the combination of different contributions:

$$\mathcal{L} = L_{\text{cls}} + L_{\text{box\_reg}} + L_{\text{rpn\_cls}} + L_{\text{rpn\_loc}} + L_{\text{mask}} \tag{2.2}$$

where L_{cls} is the loss in the classification head, $L_{\text{box\_reg}}$ is the loss in bounding-box regression head, $L_{\text{rpn\_cls}}$ is the classification loss in the RPN, $L_{\text{rpn\_loc}}$ is the localization loss in the RPN, and L_{mask} is the loss in segmentation head. Furher details regarding the loss equations can be found in the original Mask R-CNN paper [166].

Performance metrics and statistical analysis

As a primary endpoint, the segmentation performance was evaluated using the DSC, a statistical validation metric based on the spatial overlap between the predicted A_{mask} and the ground-truth

A_{gt} segmentation:

$$DSC = \frac{2 \times |A_{\text{gt}} \cap A_{\text{mask}}|}{|A_{\text{gt}}| + |A_{\text{mask}}|} \tag{2.3}$$

DSC can assume values in a range from 0, indicating no overlap, to 1, indicating complete overlap.

Furthermore, outcomes were also evaluated using the following spatial overlap-based metrics: Acc, represents the percent of pixels in the image that are correctly classified.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

where TP, TN, FP, FN denote the true positives, true negatives, false positives and false negatives, respectively.

Rec, also known as Sen or True Positive Rate, defines the portion of positive pixels in the ground truth which are also identified as positive in the predicted segmentation.

$$Rec = \frac{TP}{TP + FN} \tag{2.5}$$

Spec, or True Negative Rate, measures the portion of negative pixels (background) in the ground truth, that are also identified as negative in the predicted segmentation.

$$Spec = \frac{TN}{TN + FP} \tag{2.6}$$

Prec, or Positive Predictive Value, measures how accurate the predictions are, i.e. the percentage of correct predictions.

$$Prec = \frac{TP}{TP + FP} \tag{2.7}$$

F1-score is a balance between Prec and Rec, also known as harmonic mean.

$$F1\text{-}score = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}} \tag{2.8}$$

IoU, also referred to as Jaccard index, represents the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.9}$$

Mean Average Precision (mAP), which represents the average of the area under the Prec-Rec curve, was also computed.

Outcomes were compared between the different subsites analyzed using non-parametric statistics. The Kruskal-Wallis H-test was used for the overall comparison, and the Mann-Whitney U rank test for pair comparisons.

2.4.3 Results

Overall performance

The proposed model demonstrated the ability to correctly predict 39 out of 51 test images (76.5%).

Table 2.3: Summary of the diagnostic performance according to the different metrics evaluated: Dice Similarity Coefficient (DSC), Accuracy (Acc), Specificity (Spec), Precision (Prec), Recall (Rec), Intersection over Union (IoU), F1-score. Values are reported as mean ± standard deviation.

Metric	Overall	Larynx/hypopharynx	Oral cavity	Oropharynx
DSC	0.79 ± 0.23	0.90 ± 0.05	0.60 ± 0.26	0.80 ± 0.30
Acc	0.91 ± 0.12	0.98 ± 0.01	0.79 ± 0.13	0.92 ± 0.14
Spec	0.93 ± 0.12	0.98 ± 0.01	0.86 ± 0.16	0.92 ± 0.15
Prec	0.85 ± 0.24	0.94 ± 0.06	0.73 ± 0.32	0.79 ± 0.36
Rec	0.86 ± 0.22	0.91 ± 0.08	0.73 ± 0.33	0.95 ± 0.04
IoU	0.73 ± 0.27	0.87 ± 0.09	0.49 ± 0.30	0.76 ± 0.14
F1-score	0.80 ± 0.23	0.92 ± 0.05	0.61 ± 0.27	0.81 ± 0.31

The average DSC score was 0.79 ± 0.22 (range 0.26-0.97). Overall and site-specific performance metrics are summarized in Table 2.3 and Fig. 2.16, while samples of the segmentation results are presented in Fig. 2.17.

Laryngeal/hypopharyngeal lesions

The total number of laryngeal and hypopharyngeal lesions in the test set were 27 (52.9% of the test dataset). Out of that number, our algorithm correctly predicted 21 lesions (77.8%). The mean DSC score was 0.90 ± 0.05, the first quartile was 0.90 and the third quartile 0.94 (Table 2.3).

Oral lesions

The oral lesion frames comprised in the test set were 15 (29.4% of the total). The algorithm performed a correct prediction in 13 cases (86.7%). The mean DSC score was 0.60 ± 0.26, the first quartile was 0.34 and the third quartile 0.84 (Table 2.3).

Oropharyngeal lesions

In the test set, the oropharyngeal lesions were 9 of 51 images (17.6%). The algorithm correctly predicted 5 images (55.5%). The mean value of DSC score was 0.81 ± 0.30, the first quartile was 0.92 and the third quartile 0.95 (Table 2.3).

Comparison between three different UADT sites

Results for each site are summarised in Table 2.3. The overall diagnostic performance, defined by the DSC score, was significantly different between the different sites ($p = 0.002$). Pairwise analysis showed that the difference was related to significantly inferior results in the oral cavity when compared with larynx/hypopharynx ($p < 0.001$). Diagnostic results proved to be significantly correlated with the site analyzed also considering other performance metrics: Acc ($p < 0.001$), Spec ($p = 0.02$), IoU ($p = 0.002$), and F1-score ($p = 0.002$). As above, this difference is related to inferior results in the oral cavity vs larynx/hypopharynx. However, when considering Acc, it is also possible to evidence a significant difference between the oral cavity and oropharynx ($p = 0.03$).

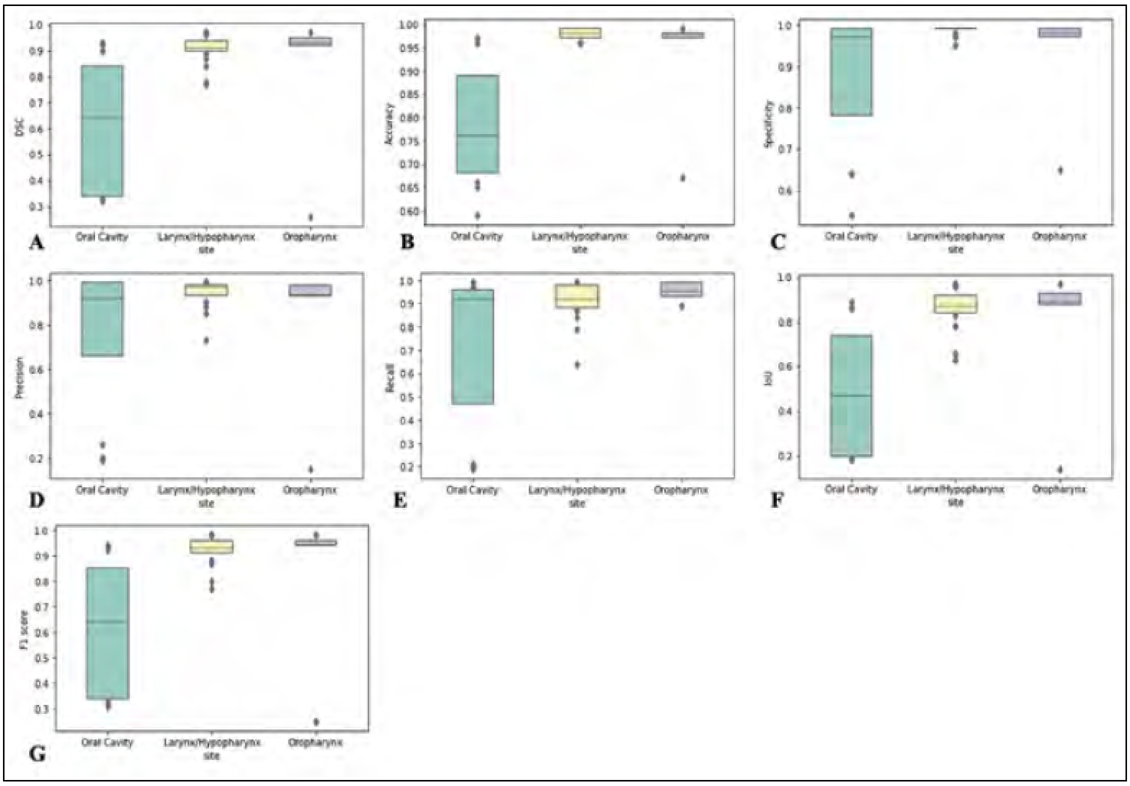


Figure 2.16: Box plots detailing the diagnostic accuracy of the algorithm in different sites according to various metrics. (A) Dice Similarity Coefficient (DSC); (B) Accuracy (Acc); (C) Specificity (Spec); (D) Precision (Prec); (E) Recall (Rec); (F) Intersection over Union (IoU); (G) F1-score.

2.4.4 Discussion

This study, for the first time in literature, evaluated the specific task of instance segmentation in clinical endoscopy for head and neck SCC. The analysis included three sites of the UADT to allow comparison of the algorithm's diagnostic performance in different anatomical areas. The algorithm was able to identify and segment the lesion in 76.5% of cases, and showed remarkable diagnostic accuracy, especially in consideration of the complex task to be performed. Interestingly, results were significantly inferior in the oral cavity, where all outcome measures underperformed when compared with the larynx/hypopharynx and, in some cases (i.e., Acc), oropharynx. This is in line with what was previously observed by Piazza et al. [95] when applying bioendoscopic tools such as NBI. This result is possibly related to the wide array of epithelial subtypes observed in the oral cavity, adjunctive limits specifically correlated with oral examination (e.g., presence of light artifacts), and confounding factors (e.g., tongue blade, teeth, or dentures) that the DL algorithm must learn to take into account. Instance segmentation represents the ultimate step in video analysis since it allows at the same time detection, classification, and segmentation of multiple elements in each single frame, which is possible thanks to the integration of different analytic components in the same general algorithm. This approach is particularly suited to the context of UADT endoscopy since different alterations (e.g., concomitant inflammatory or benign lesions) can be frequently encountered in the field of view together with the target lesion, and due to the fact that patients with head and neck SCC can develop distinct islands of neoplastic or dysplastic mucosa (i.e., field of cancerization) that might involve various portions of the videoframe, even without continuity.

Recent CNN-based methods have demonstrated remarkable results in the segmentation of the

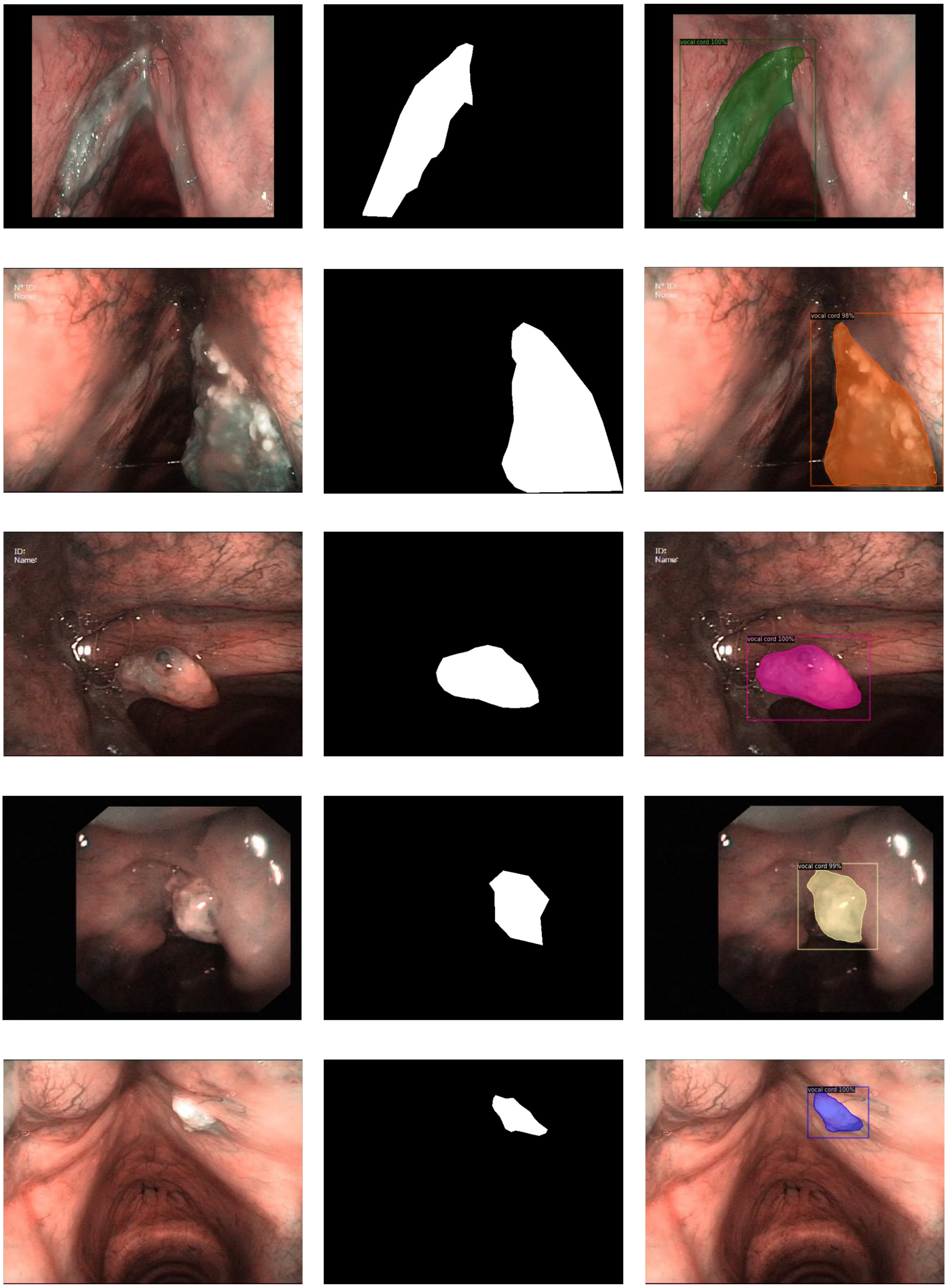


Figure 2.17: Visual samples of the segmentation results. From left to right: raw endoscopic frames, ground truth annotation, and predictions obtained with the proposed method.

UADT and proved to be well-suited for such a complex task. Laves et al. [113] first demonstrated that a weighted average ensemble network of U-Net and ErfNet were the best suited for laryngeal segmentation of intra-operative images under direct laryngoscopy, with a mean IoU of 84.7%. However, different authors subsequently strived toward the development of diagnostic algorithms that could be applied in real-time in office-based and intra-operative endoscopy. Paderno et al. [114] explored the use of fully CNNs for real-time segmentation of SCC in the oral cavity and oropharynx. In this work, different architectures were compared detailing their diagnostic performance and inference time, demonstrating the possibility of achieving real-time segmentation. In accordance with previous findings in the literature, the present study confirms that the oral cavity may have inferior diagnostic results due to the high variability of subsites when compared with other areas of the UADT (i.e., oropharynx, larynx, and hypopharynx). When dealing with normal laryngeal anatomy, Fehling et al. [115] explored the possibility of achieving a fully automated segmentation of the glottic area using a CNN in high-speed laryngeal videos. The algorithm obtained a DSC over 0.85 for all subsites analyzed. Finally, Li et al. [116] proposed a method to segment nasopharyngeal malignancies in endoscopic images based on DL, reaching an Acc of 88.0%. However, progressive advances in automatic segmentation of the UADT can be observed thanks to a recent article by Azam et al.[161], in which Seg-MENT, a novel CNN-based segmentation model, outperformed previously published results on the external validation cohorts. The model was initially trained on white light and NBI endoscopic frames of laryngeal SCC, but also showed to be effective in the segmentation of independent frames of oral and oropharyngeal cancer. The authors stated that the model demonstrated potential for improved detection of early tumors, more precise biopsies, and better selection of resection margins.

In general, results of automatic segmentation are inferior to those obtained in more straightforward tasks such as frame classification [102, 105, 108] or lesion detection [109, 112] since a more in-depth conceptual model of UADT lesions is required to allow accurate definition of margins. However, semantic segmentation is a key objective when striving towards more complex tasks involving computer vision and human-machine interaction. In fact, other than providing a purely diagnostic tool, a comprehensive understanding of all UADT alterations and suspicious lesions may grant significant aid in intra-operative management. This is even more true when considering instance segmentation, which epitomizes in itself all the needs and requirements of the visual examination of endoscopic images, allowing a full automatic understanding of complex endoscopic scenarios, even those involving more than one lesion and/or more than one pathology.

Potential issues have been addressed to limit biases related to the analysis technique:

- Patients (and their related frames) in the training, validation, and test sets have been distinguished into separated groups to avoid overfitting;
- Frames were annotated and reviewed by 3 experts to limit subjective errors;
- Frame selection and data augmentation were performed to reduce the impact of artifacts or technical biases.

However, intrinsic limits should be acknowledged. In particular, the gold standard over which the algorithm has been trained (i.e., the “ground truth”) is represented by an expert opinion of the tumor margins and not by the histo-pathological definition per se. In fact, as of today, it is not technically possible to provide a direct in situ, in vivo morphologic correlation between endoscopic images and their histopathological specimen.

Chapter 3

Artificial intelligence-driven applications in magnetic resonance imaging archives

MRI is a non-invasive medical imaging technique that uses non-ionizing and harmless radiation to offer a detailed 3D visualization of internal anatomical structures [169]. It operates on the principle of utilizing strong magnetic fields and radiofrequency waves to generate images of soft tissues composing the body. The procedure begins with the placement of the body within a powerful magnetic field, initially with the magnets switched off, causing the water molecules in the body to reach their equilibrium position; subsequently, the magnetic field is activated, prompting the water molecules to align with the magnetic field's direction [170]. To stimulate the protons within the body, a powerful radiofrequency energy pulse is applied in alignment with the magnetic field, causing them to spin against the magnetic force; then, upon deactivation of the radiofrequency energy pulse, the water molecules return to their equilibrium position and realign with the magnetic field. During this realignment, the water molecules emit radiofrequency energy, detectable by the scanner, and converted into visible images [171]. The quantity of radiofrequency energy emitted depends on tissue structure and its intensity can be adjusted through scanner parameter variation, enabling the production of multiple modality images. Critical factors that determine MRI images are:

- TE time (time to echo): the time between the delivery of the radiofrequency pulse and the receipt of the echo signal [170];
- TR time (repetition time): the amount of time between successive pulse sequences applied to the same slice [170].

The MRI machine's ability to capture multiple images from various angles with different contrasts and physical properties classifies it as a versatile multiple modality imaging tool, widely applied in various medical contexts, including neurological examinations, musculoskeletal studies, and cardiac imaging [169].

DL techniques have the potential to significantly improve MRI analysis by automating complex tasks such as anatomy segmentation, pathology detection, and disease classification. The high-dimensional and multichannel nature of MRI data, along with potential variations introduced by hardware, artifacts, and patient anatomies [169], make DL a promising tool for interpreting heterogeneous MRI images. Some DL-based MRI solutions have gained Food and Drug Administration

(FDA) approval for tasks like image reconstruction, enhancement, segmentation, and classification [172]. Nevertheless, research on the advancement of this technology is ongoing, and a notable gap remains in its implementation in standard care. Overall, the different application areas of DL in MRI can be broadly categorized into model-free image synthesis, model-based reconstructions, and either pixel- or image-level classification task [173].

The first category revolves around model-free image synthesis, a transformative approach that enables the synthesis of one MR image contrast or artifact status into another. For instance, many common image synthesis tasks include MR image super-resolution [174], image denoising [175], artifact reduction [176], and contrast synthesis [177].

The second category, model-based reconstructions, leverages DL in combination with k-space data. These models embed the fundamental principles of Fourier encoding and data consistency into iterative DL algorithms. This integration enhances the reconstruction of MRI images from raw data, leading to significant improvements in image quality and reducing acquisition time [178, 179]. By incorporating physics-based constraints, DL-assisted model-based reconstructions offer remarkable potential for faster and more precise MRI scans across various clinical applications.

Lastly, DL in MRI finds application in pixel- or image-level classification tasks. This category involves the development of DL methods to determine variables pertinent to the diagnostic status of the subject from either 2D MRI slices, 3D volumes, or multiple 3D sequences. For example, DL models can be trained to identify and classify lesions, tumors, or anatomical structures within MRI scans [169]. Additionally, DL-based segmentation techniques can be employed to delineate specific tissues or regions of interest, streamlining image analysis and enabling quantitative measurements. These advancements in classification and segmentation tasks enhance diagnostic accuracy and streamline the interpretation of MRI data, ultimately benefiting patient care and clinical decision-making [173].

It is within this last application area that the method described in this Chapter is situated. A DL method specific for orthopedics and spinal care is presented: a new self-supervised Domain Adaptation (DA) approach for the automated segmentation of Intervertebral Disks (IVDs) from MRI images to facilitate treatment planning.

Among the various imaging techniques that can be used to visualize and study IVDs, MRI stands out as the preferred choice due to its ability to generate high-quality images with a high level of detail, and without ionizing radiation. IVD segmentation from these imaging studies holds significant importance in various medical applications associated with the spine, from diagnosis to treatment, as well as image-guided interventions. However, MRI still presents intrinsic challenges, such as the variability in imaging techniques across different machines and the lack of standard protocols for disk segmentation. This work demonstrated its ability to address the problem of IVD segmentation in MRI enhanced by DA, paving the way for future research in the field. From a clinical perspective, the proposed framework proved to be a valuable tool to support the clinical routine, increasing the diagnosis accuracy. Furthermore, the integration of DA introduces a multicentric aspect facilitating cooperation and ensuring consistent and accurate segmentation across different MRI scanners and settings.

A journal paper on this work is currently under review at the International Journal for Computer Assisted Radiology and Surgery as: Fiorentino, M.C. & Villani F.P., Benito Herce, R., González Ballester, M.A., Mancini, A., and López-Linares Román, K. (2023), "An Intensity-based

Self-supervised Domain Adaptation Method for Intervertebral Disc Segmentation in Magnetic Resonance Imaging".

3.1 Introduction to a self-supervised domain adaptation approach for intervertebral disc segmentation

IVDs are intricate structures composed of fibrocartilage, located between the vertebrae of the spine. They play a crucial role in enabling the spine to flex, twist, and distribute compressive forces evenly across the adjacent vertebral bodies [180]. Multiple medical imaging techniques are used for the visualization of IVD, including CT, X-ray, thermal imaging, and MRI [181]. Among these techniques, MRI is the preferred choice due to its ability to generate high-quality images with a high level of detail, all without the need for ionizing radiation, which can be detrimental to one's health [182].

The segmentation of IVD from these imaging studies holds significant importance in various medical applications associated with the spine, encompassing the diagnosis and treatment of spinal conditions and diseases, as well as image-guided interventions [183]. An example of this segmentation is shown in Fig. 3.1. For IVD segmentation tasks, DL methods, particularly CNNs, are commonly used and have demonstrated reliable results [184]. Most research on IVD segmentation relies on supervised DL approaches, mainly based on encoder-decoder architectures [185–189], fully convolutional models [190], mixed supervised methods [191–193], or more recently, on semi-supervised models for simultaneous segmentation of vertebral bodies and IVD [194]. These researches often exhibit performance limitations when applied to data from different acquisition devices and protocols. This is because these methods are typically trained on a specific dataset of images from a specific device or patient population. As a result, they present difficulties when applied to images from sources different from the training data [195]. This aspect, known as domain shift, becomes especially evident in the case of MRI, where different scanners and acquisition modalities produce very different images in terms of intensity distribution. Intensity variations might arise when using different scanners due to factors like drift in scanner signal-to-noise ratio over time, gradient non-linearity, or due to changes in scanning protocol parameters (flip angle, echo, or repetition time) [196]. Furthermore, images acquired with different MR modalities, such as T1 and T2-weighted, result in very different contrast and brightness characteristics.

In different fields of medical image analysis, various DA techniques have been used to address the challenges associated with transferring knowledge from a source domain to an unlabelled target domain. These techniques include adversarial learning, which aligns feature distributions [197, 198], self-ensembling methods that generate pseudo-labels for unlabeled target domain samples using trained model predictions [199–203], cycle consistency models that synthesize target images from source images [204–206], and DA methods based on variational autoencoders [207].

To enhance the transferability of models to target domains, the inclusion of auxiliary tasks has recently become a prominent strategy, in the field of computer vision, to develop domain-invariant representations [208]. By integrating these auxiliary tasks, models have the capability to acquire extra valuable features, leading to enhanced performance during DA. This approach offers the advantage of being easy to train, making the overall process more manageable and efficient. This integration opens up new avenues for effectively addressing DA challenges in medical image analysis,

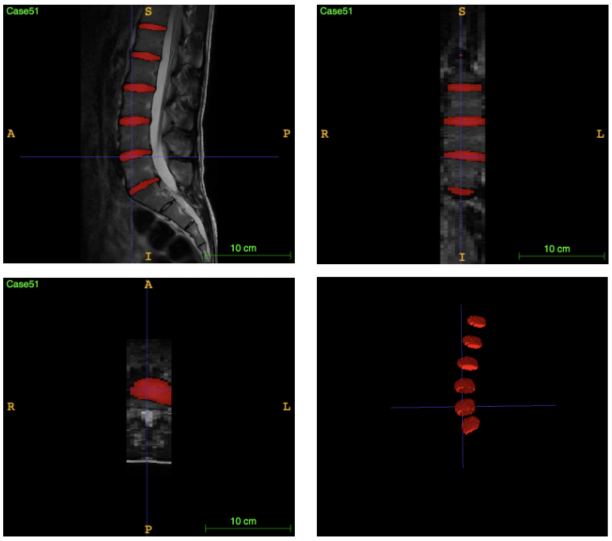


Figure 3.1: The 3D intravertebral disc segmentation enables precise quantification and analysis, facilitating the evaluation of various spine-related disorders (e.g. herniated discs, degenerative disc disease, and spinal stenosis etc).

in which traditionally data challenges limited model generalization across different institutions and patient populations. Currently, in the medical field, only few studies have investigated the use of auxiliary tasks for segmentation. A structure-driven DA approach for unsupervised cross-modality cardiac segmentation is proposed in [209]. A set of 3D landmarks serves as representative points that embody the common anatomical structure of the heart across various imaging modalities (CT and MRI). The model learns to predict the positions of these landmarks, facilitating the identification and use of shared structural information. Cardiac structures segmentation from CT and MR volumes is also explored in [210], which makes use of an edge generation auxiliary task to support the primary segmentation task in the target domain. To cope with domain shift, they employ hierarchical low-level adversarial learning to encourage the suppression of informative features in a hierarchical manner. [211] focus on unsupervised DA for abdominal multi-organ segmentation on CT scans, leveraging the organ location information. A jigsaw puzzle auxiliary task is devised, where a CT scan is reconstructed from shuffled patches. Additionally, a super-resolution network is used to standardize images from multiple domains. The auxiliary and super-resolution tasks are trained alongside the organ segmentation task to enhance overall performance.

However, to the best of our knowledge, this approach has not yet been investigated in the context of IVD segmentation.

The effectiveness of the DA based on the inclusion of auxiliary tasks strongly depends on the optimal design of the pretext task, which can present challenges, such as the domain shift between the pretext task and the final segmentation domains [212]. The characteristics and variations present in the initial unlabeled dataset used for the pretext task must, in fact, closely align with those of the final segmentation task domain. This facilitates effective knowledge transfer from the pretext task to enhance the accuracy and robustness of the final segmentation.

Inspiration is drawn from recent research in the field of fashion compatibility [213, 214], recognizing the significance of leveraging color and texture as valuable factors for understanding and categorizing visual data. This research, in fact, has introduced the application of color and texture pretext tasks as a strategy to acquire discriminative features of the data, while disregarding shape information. Color features in natural images translate into intensity features in medical imaging, as shown by Galdran et al. [215] which recently reported a similar approach for classification of histopathological images and out-of-distribution detection. This aspect is of particular interest for MR images, for which variations in hardware and software create non-standard tissue intensities. Nevertheless, it is worth noting that this specific type of pretext task has yet to be explored in the field of medical image segmentation, particularly in the context of self-supervised DA.

Thus, guided by these considerations, the contributions of this work can be summarized as follows:

1. First attempt at exploring self-supervised DA for the segmentation of IVD in MRI.
2. First-ever work that leverages self-supervised learning specifically introducing intensity-pretexts, for DA in the domain of medical image segmentation.

All the experiments are performed using publicly available datasets to promote comparisons with the presented methods.

3.2 Materials and methods

This section introduces the proposed framework along with the datasets used, and the training settings.

3.2.1 Dataset description

Three datasets were used to leverage DA for IVD segmentation: one as the source domain (S), while the other two smaller datasets were designated as the target domains ($T1$ and $T2$). The three datasets are publicly available and were chosen based on their diverse characteristics, which provide a broader overview of the anatomical region under consideration. These datasets were collected at different medical centers and using different MRI scanners, thus they differ in terms of the patient population and MRI parameters used, further increasing the heterogeneity of the data.

S: The dataset S was publicly released by [216], it was obtained from a single hospital in China and includes T2-weighted MR volumetric images of 215 subjects, acquired with a 3.0 Tesla MRI scanner (Ingenia, Philips, Amsterdam, Nederlands). Among these subjects, there are 6 individuals without any spinal abnormalities, serving as the control group. The remaining subjects have various spinal conditions, with 177 patients diagnosed with vertebrae degeneration, 204 patients with IVD degeneration, 21 patients with lumbar spondylolisthesis, 91 patients with spinal canal stenosis, 22 patients with schmorl's nodes, and 53 patients with vertebral endplate osteochondritis. Patients in the dataset may be simultaneously affected by multiple spinal diseases or disorders. For further details refer to [216]. This dataset has manual delineation of the vertebrae and IVD performed by a junior expert, and subsequently reviewed and corrected by a senior expert. The in-plane resolutions range from 512×512 to 1024×1024 pixels, with a pixel spacing ranging from 0.30mm to 0.59mm (average of 0.35mm), slice thicknesses ranging from 4.40mm to 5.50mm (average of 4.42mm), and number of slices ranging from 12 to 18. For training, validating and testing the model, the dataset was split into three sets including 172, 19, and 19 volumes, respectively.

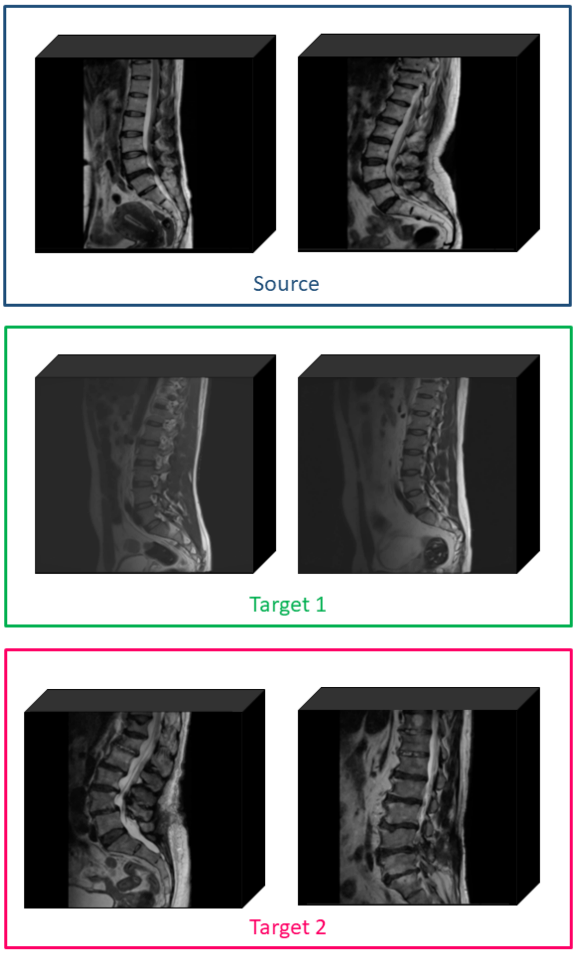


Figure 3.2: Intervertebral discs in magnetic resonance imaging show a wide range of variations in their appearance, including signal intensity, shape, and texture. Additionally, factors like aging, degeneration, and pathology contribute to the diverse appearances of the discs. The image showcases the distinct characteristics found within datasets, emphasizing the impact of spinal pathology on the integrity of the spine with diverse contrast levels and illuminations. Target 2, which includes images acquired with various scanners and pathological patients, exhibits additional complexity compared to the other two datasets.

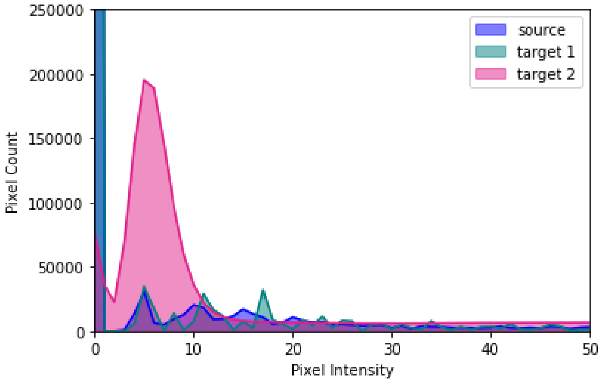


Figure 3.3: Intensity histograms of the three datasets. The Source dataset exhibits a pronounced peak near zero, indicating a predominant concentration of low-intensity values; Target 1 presents a flatter distribution, indicating well-contrasted images; Target 2 displays a bell-shaped distribution centered around pixel intensity of 5. For visualization purposes, the plot is truncated at an intensity value of 50 and a pixel count of 250000.

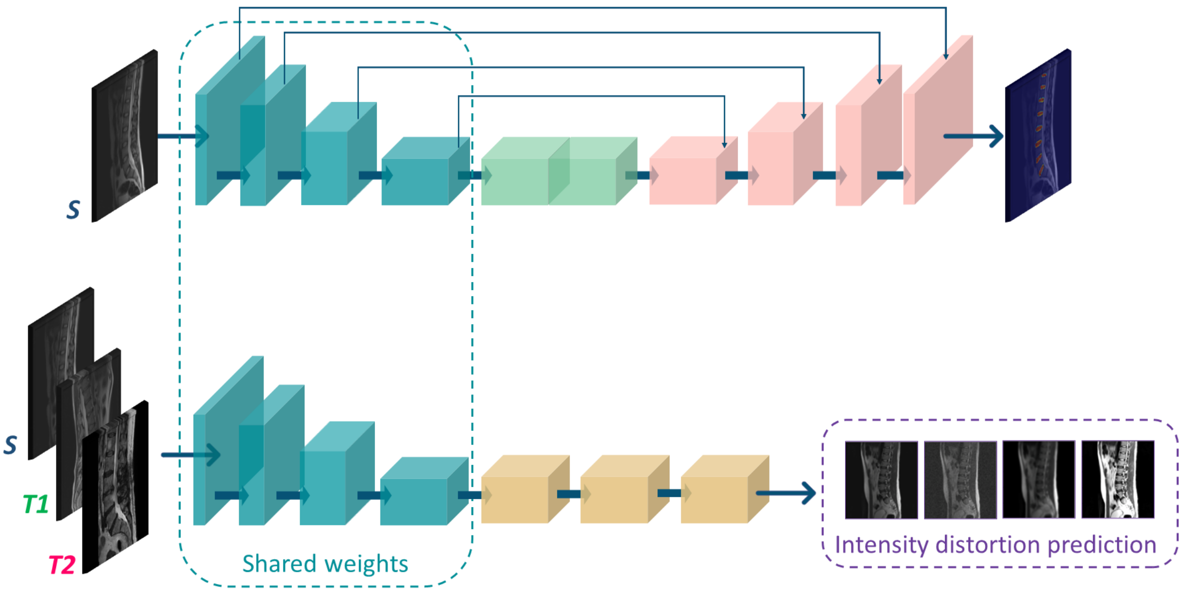


Figure 3.4: The proposed framework for self-supervised domain adaptation which focuses on learning domain-invariant feature intensity representation. This is achieved by incorporating a pretext learning task that automatically generates labels from images of both the source and the two target domains. The pretext task and the main task, which is intervertebral disc segmentation, are simultaneously trained using multi-task learning.

$T1$: The dataset $T1$ was released by [217], it consists of T2-weighted MR images from 23 patients, acquired with a 1.5 Tesla MRI scanner of Siemens (Siemens Healthcare, Erlangen, Germany). All the images are sampled to the same sizes of 39 × 305 × 305 voxels. The pixel spacing of all the images is $2 \times 1.25 \times 1.25$ mm^3. The authors do not provide further information about the medical condition of the patients. No IVD labels are available for these data. The dataset was split into three sets for training, validation, and testing, each set including 14, 4, and 5 volumes respectively. The testing set was meticulously labeled by three experienced operators.

$T2$: The dataset $T2$ was publicly released by [218], it consists of 30 MR volumes from 29 patients with different medical conditions including chronic low back pain, known malignancy with spinal metastases, spondylodiscitis or other spinal inflammatory or infectious diseases, and spinal fracture. Various scanner models from different vendors were used to collect the dataset, which is composed of 70.9% of images derived from Philips scanners (Achieva, Ingenia, and Elition), 27.3% derived from Siemens scanners (Avanto, Verio, Espree, Symphony, Amira, Aera, and Magnetom), and 1.8% (a single MR volume) taken from a GE scanner (Signa). The in-plane resolutions range from 384 × 336 to 1200 × 1200 pixels, with a pixel spacing ranging from 0.24 × 0.24 × 3.00 mm to 0.91 × 0.91 × 4.80 mm (average of 0.47 × 0.47 × 3.46mm), slice thicknesses ranging from 3.00mm to 4.80mm (average of 3.46mm), and number of slices ranging from 12 to 27. MR scans from 19 patients were used to train the model, 5 to validate it, and 6 for testing. Also in this case, three operators with significant experience carefully assigned labels to the testing set.

All voxels belonging to IVD were labeled as 1 in all three datasets while all the other voxels were set to 0. To ensure reproducibility and facilitate comparisons, Table 3.1 presents the subsets used for model testing in each of the three datasets, along with the corresponding scanners involved. For

Table 3.1: Test sets from the three datasets. The magnetic resonance scanners used to collect the images are reported, together with the identification number (ID) of the images. S refers to the source dataset, $T1$ to the target 1 dataset, and $T2$ to the target 2 dataset.

Dataset	Scanner	ID
S	Philips Ingenia	3, 11, 15, 21, 49, 51, 60, 63, 74, 86, 114, 117, 119, 125, 131, 135, 151, 183, 192
$T1$	Siemens	1, 9, 10, 16, 18
$T2$	Philips Achieva	23
	Philips Elition	24, 32
	Siemens	33, 36
	GE Signa	38

Algorithm 1: Domain adaptation method pseudo-code.

```
Input  : labeled source domain images $(X_s, Y_s)$ and unlabeled target domains images $(X_t)$
Output: trained $(e(\cdot))$, trained $(d(\cdot))$
Sampling for pretext task: $(X_s, Y'_s)$ and $(X_t, Y_t)$
Generation of 4 different samples from $X_s$ and $X_t$, labeled as $Y_t$,$Y'_s$, according to 3 different
  intensity transformations + no transformation applied
while $epoch \leftarrow 0$ to $epochs$ do
    load target + source mini-batch $(x_{ti}, y'_{ti})$, $(x_{si}, y'_{si})$;
    forward pass and compute $L_p$;
    back-propagate $L_p$ gradients by $(e(\cdot))$
    update weights of $(e(\cdot))$
    load source mini-batch $(x_s i, y_s i)$
    forward pass and compute $L_s eg$
    back-propagate $L_{seg}$ gradients by $(e(\cdot))$ and $(d(\cdot))$
    update weights of $(e(\cdot))$ and $(d(\cdot))$
end
return $e(\cdot)$, $d(\cdot)$
```

each dataset, the images included in the test sets were selected to ensure no patient overlap between the train and test sets. Moreover, to guarantee a robust evaluation of the model's adaptability to a wide range of acquisition contexts, for dataset $T2$, which includes images acquired with various scanner models, the test set was selected to contain images acquired with scanner models not present in the dataset S. Samples of the testing set images are shown in Fig. 3.2.

A visual perspective of the intensity distribution for the MR images across the datasets is provided in Fig. 3.3. By juxtaposing the histograms, it is possible to discern the variations and similarities in intensity profiles, a crucial facet for DA strategies.

3.2.2 Proposed method

Here, is presented the proposed approach consisting of a dual-task model based on a CNN architecture. The model primarily focuses on IVD segmentation as its main task, while simultaneously incorporating a contrastive pretext learning task. The pretext task automatically generates labels from images gathered from both the primary domain and two additional target domains, classifying

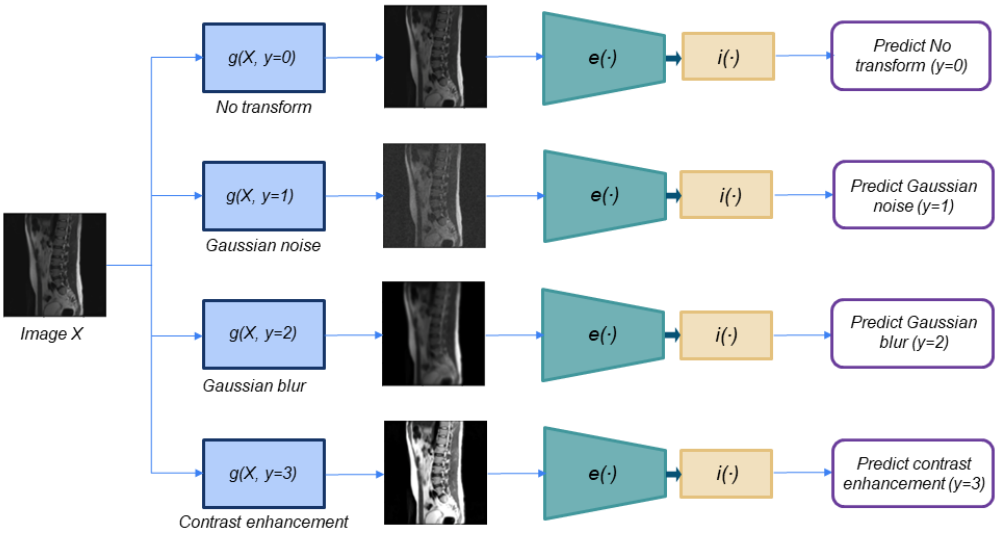


Image X
g(X, y=0)
No transform
g(X, y=1)
Gaussian noise
g(X, y=2)
Gaussian blur
g(X, y=3)
Contrast enhancement
e(·)
i(·)
Predict No transform (y=0)
Predict Gaussian noise (y=1)
Predict Gaussian blur (y=2)
Predict contrast enhancement (y=3)

Figure 3.5: Proposed pretext task that automatically generates labels from images of both source and target domain.

them based on the transformations applied. The inclusion of the pretext task is key to deriving features that are invariant across different domains.

When analyzing datasets obtained from different MR devices, it is crucial to take into account scanner-dependent variation in image signal intensity. Thus, the main objective within the framework of self-supervised learning is to obtain intensity representations within the embedding space. This effort is aimed at improving the model's ability to generalize and handle intensity variations more effectively.

Fig. 3.4 shows the proposed method focused on IVD semantic segmentation as the main task. To accomplish this, an encoder network $e(\cdot)$ is employed, which serves as a feature extractor. Additionally, a decoder network $d(\cdot)$ is used to recover spatial information and generate accurate IVD segmentations. The $e(\cdot)$ is composed of four blocks, each consisting of two 3D convolutional (conv) layers with a kernel size of 3x3x3 and same-padding. Following the conv layers, a ReLU activation function and a batch normalization layer are applied. A max pooling operation with a stride of 2x2x1 is performed. At each block, the number of channels doubles, enabling the incremental learning of more complex features. The number of channels starts from 32 and progressively increases to 512. The $e(\cdot)$ also incorporates a bottleneck section which facilitates the connection to the $d(\cdot)$, as suggested in [219]. This bottleneck section consists of two additional 3D conv layers with a kernel size of 3x3x3 and same-padding. Subsequently, a ReLU activation function and a batch normalization layer are applied to further enhance the learned representations. Similarly to $e(\cdot)$, the $d(\cdot)$ function is composed of four blocks. Each block comprises two 3D conv layers with a kernel size of 3x3x3. These layers are followed by a ReLU activation function and an upsampling layer with a kernel size of 2x2x1, which reduces the number of feature channels by half. To recover the lost features resulting from downsampling in the $e(\cdot)$ path, the input of each block is concatenated with the corresponding feature maps from $e(\cdot)$. The last block consists of three 3D conv layers, with the first two being followed by a ReLU activation function, and the last one activated by softmax. The number of filters used in the conv layers starts at 256 and is halved in each subsequent block until reaching 32 filters. This CNN is trained end-to-end using labeled samples from the source domain ($S = \{X_s, Y_s\}$).

In order to learn intensity invariant features, $e(\cdot)$ is also trained to recognize intensity distortions from both target domains ($T = \{X_t, Y_t\}$) and source domain (X_s, Y_s'). Y_t and Y_s' are derived automatically by applying image intensity transformations to their respective images and labeling them based on the specific transformation applied. This is further detailed in Section 3.2.3.

The entire DA method is outlined in Algorithm 1. This illustrates that self-supervised DA encompasses the simultaneous training of models for both the pretext and the main task. During forward propagation, samples from both the source and target domains are processed by the shared encoder. Subsequently, the losses for the main task (L_{seg}) and the pretext task (Lp) are calculated, and these losses are then back-propagated and accumulated at $e(\cdot)$. By training $e(\cdot)$ with samples from all three domains, the model learns feature representations that are invariant to domain differences. In the testing phase, target domain images are inputted to $e(\cdot)$, and the resulting features are transmitted to $d(\cdot)$ of the main task, which enables to obtain predictions.

3.2.3 Pretext tasks

The intensity prediction pretext task proposed in this work is inspired by prior works from different fields [214], [213], [208]. An overview of the pretext task model is provided in Fig. 3.5. Given a set of N_t and N_s training images from $T = \{x_i^t\}_{i=0}^{N_t}$ and $S = \{x_i^s\}_{i=0}^{N_s}$, respectively, three different sets of intensity transformations, namely Gaussian noise, Gaussian blur, and contrast enhancement are applied. The intensity transformation prediction model $i(\cdot)$, takes the feature maps generated by the function $e(\cdot)$ as input and produces a probability distribution representing different intensity transformations, including the option of no intensity transformation. The $i(\cdot)$ model is composed of three blocks, each containing two 3D conv layers. These conv layers have a kernel size of 3x3x3 and use same-padding. The ReLU activation function is applied after each of them, and batch normalization is performed subsequently. The number of filters used in the conv layers starts at 256 and is halved in each subsequent block until reaching 64 filters. Furthermore, an additional 3D conv layer with a kernel size of 3x3x3, same-padding, and a number of filters denoted as C is employed to reduce the number of filters to match the number of classes in the problem. This additional layer is activated by the softmax function.

The L_p loss is defined as:

$$\begin{aligned} L_p = -\frac{1}{C * N_t} \sum_{i=1}^{N_t} \sum_{j=1}^{C} y_{ij} log(c(e(x_i^t; \theta_e); \theta_c)_{ij}) + \\ -\frac{1}{C * N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{C} y_{ij} log(c(e(x_i^s; \theta_e); \theta_c)_{ij}) \end{aligned} \tag{3.1}$$

where C=4.

3.2.4 Parameter setting

The training process involved resizing all images from datasets S, $T1$, and $T2$ to $256 \times 256 \times 18$ pixels. Both the primary task (IVD segmentation) and the pretext task were optimized using the

Adam optimizer for 100 epochs. A fixed initial learning rate of 0.001 was used, with a batch size of 1 for the main task and a batch size of 4 for the pretext task. The Dice loss was employed as L_{seg}, which is known for its sensitivity to class imbalance. This property makes it well-suited for tasks where certain classes are scarce, such as in this context where IVD are considerably smaller than the background class.

During the IVD segmentation training, on-the-fly data augmentation was performed to enhance generalization performance. The augmentation techniques included geometrical transformations (such as horizontal flipping, and random rotation in the range of ±30 degrees) and intensity transformations (such as random brightness correction). These augmentations were randomly applied at each training iteration. For the pretext task, only geometrical transformations were used, specifically random vertical flipping and random rotation in the range of ±30 degrees. These transformations aimed to improve generalization to different perspectives without affecting the task of identifying intensity transformations in the images. The best model among epochs is selected based on the lowest $L_{total} = L_{seg} + L_p$ obtained in the validation set of S.

All the analyses were performed using Tensorflow 2.x on an NVIDIA RTX 2080 TI, with a Xeon e5 CPU and 128 GB RAM.

3.2.5 Performance metrics

The performance of this end-to-end model was evaluated by calculating metrics for 3D segmentation, as outlined in [220]. Hence, overlap-based metrics were computed on the testing datasets of S, $T1$, and $T2$ such as the DSC, Sen, and Spec (or Rec) as defined in Sec. 2.4.2, and in particular in Eq. 2.3, Eq. 2.5, and Eq. 2.6, respectively.

Furthermore, the Hausdorff Distance (HD), a distance-based metric, is employed as an additional measure for assessing boundary delineation. The HD is defined as follows:

$$HD(X,Z) = max(h(X,Z), h(X,Z)) \tag{3.2}$$

where

$$h(X,Z) = \underset{x \in X}{max} \underset{z \in Z}{max} ||x - z|| \tag{3.3}$$

3.2.6 Baseline and domain adaptation comparison

First, the stage is set by evaluating the proposed strategy against the baseline model, namely the U-Net model trained only on the S dataset. For this comparison, a further experiment is conducted to investigate the feasibility of segmenting individual IVD within the spine, in order to capture the differences in the spatial configuration and morphology of each IVD. To achieve this, a Principal Component Analysis (PCA) approach is employed to divide the overall IVD segmentation into separate individual discs segmentation.

Next, a comprehensive analysis comparing different training data configurations is conducted to investigate the impact of introducing different domains into the pretext task:

1. Dual-task model, trained by applying the pretext task exclusively on $T1$ (*t1-int*).
2. Dual-task model, trained by applying the pretext task exclusively on $T2$ (*t2-int*).

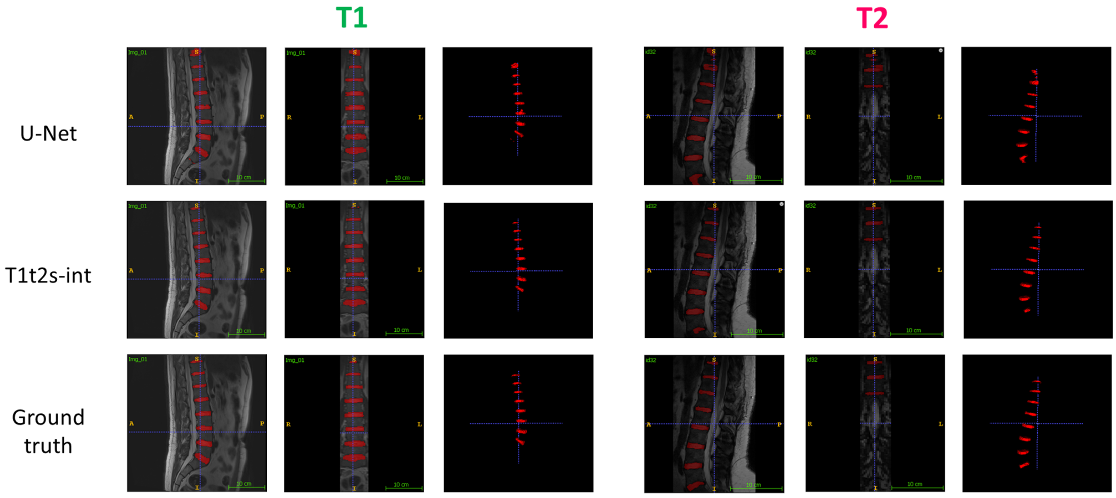


Figure 3.6: The qualitative results of U-Net and *t1t2s-int* training strategies on two random test images from target 1 and target 2 reveal notable differences. Upon examining the images, it becomes apparent that the U-Net approach yields sub-optimal disc segmentation, particularly for the discs located in the outermost regions of the image.

3. Dual-task model, trained by applying the pretext task on $T1$ and $T2$ datasets (*t1t2-int*).
4. Dual-task model, trained by applying the pretext task on both $T1$ and $T2$ and on S datasets (*t1t2s-int*).

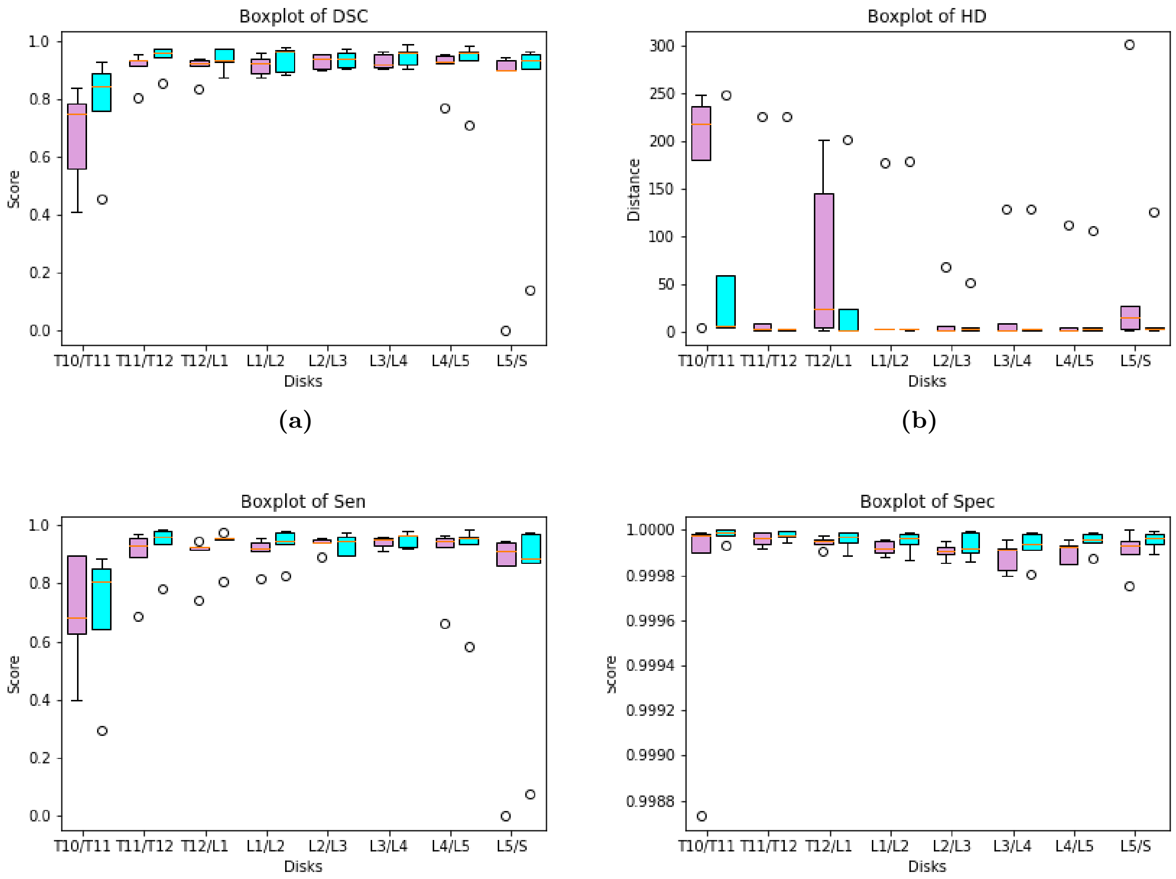
By evaluating these configurations, the aim was to assess the contributions and relative importance of individual datasets within the dual-task framework. This set of analyses enabled the understanding of the effectiveness of incorporating the pretext task across different datasets, thereby shedding light on the benefits of using dual-task learning in the proposed approach.

3.2.7 Pretext task adequacy analysis

A further experiment was conducted to examine the influence of different pretext tasks on the dual-task approach. Here, the proposed pretext task of intensity prediction (*t1t2s-int*) was compared with a more traditional pretext task, which involved predicting the rotation angle of the images. For the rotation pretext task (*t1t2s-rot*), the images were randomly rotated by 0, 90, 180, or 270 degrees, and the model was trained to classify the amount of rotation. The same model used in Sec. 3.2.3 was employed for this experiment. Similar to the previous configuration, the rotation pretext was applied to $T1$, $T2$, and S datasets. Comparing the performance of these two pretext tasks aimed to evaluate their impact on the overall dual-task framework and gain insights into the effectiveness of the proposed intensity prediction task in contrast to the traditional rotation task.

3.3 Results

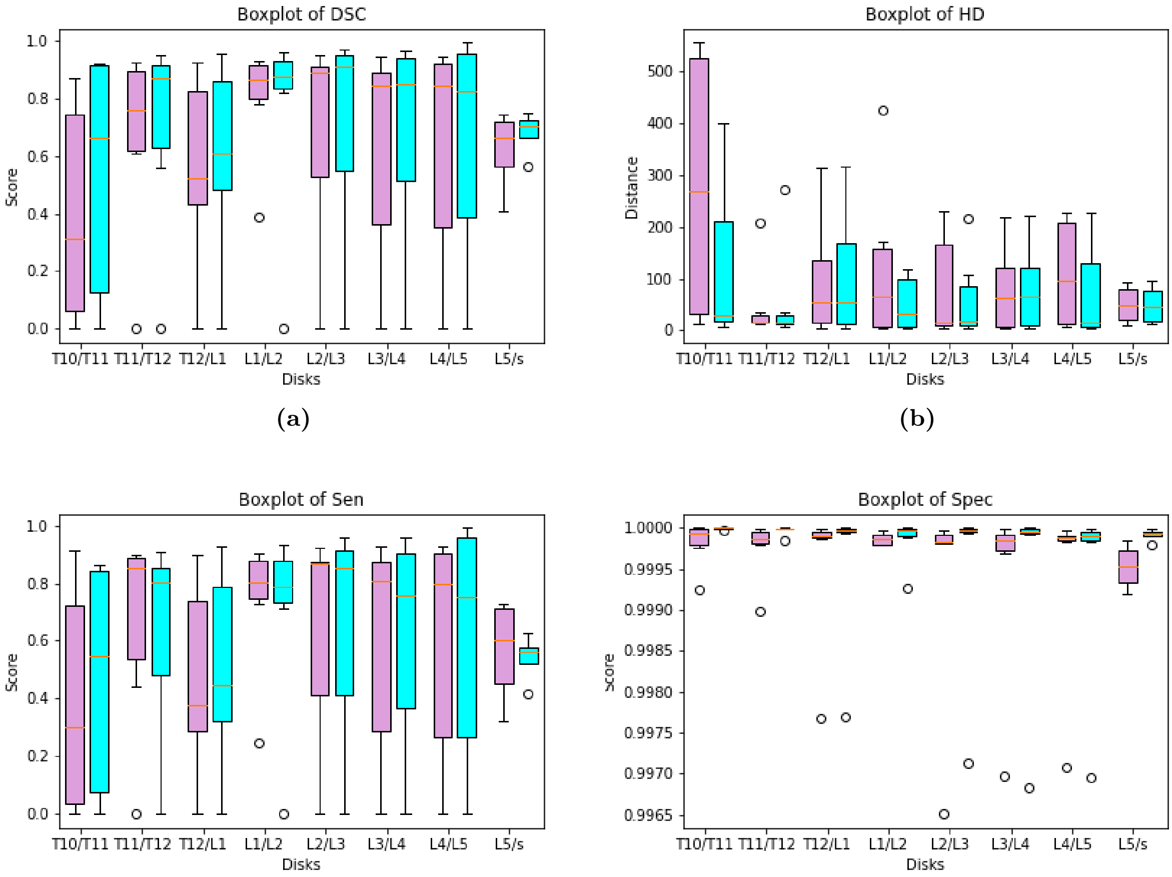
Results of the performance metrics calculated on the baseline model (i.e. U-Net trained without pretext task), and on the dual task model obtained using different training data and pretext tasks (*t1-int*, *t2-int*, *t1t2-int*, *t1t2s-rot*, and *t1t2s-int*), are presented in Table 3.2.



(a) **(b)**

(c) **(d)**

Figure 3.7: Boxplots of Dice Similarity Coefficient (DSC) (%) (a), Hausdorff distance (HD) (pixels) (b), Sensitivity (Sen) (%) (c), and Specificity (Spec) (%) (d) are presented for both U-Net (plum) and *t1t2s-int* (cyan) models, specifically for each intervertebral disc in T1. Intervertebral discs are named according to their anatomical position between vertebrae, where T, L, and S denote thoracic, lumbar, and sacral vertebrae respectively.



(a) **(b)**

(c) **(d)**

Figure 3.8: Boxplots of Dice Similarity Coefficient (DSC) (%) (a), Hausdorff distance (HD) (pixels) (b), Sensitivity (Sen) (%) (c), and Specificity (Spec) (%) (d) are presented for both U-Net (plum) and *t1t2s-int* (cyan) models, specifically for each intervertebral disc in T2. Intervertebral discs are named according to their anatomical position between vertebrae, where T, L, and S denote thoracic, lumbar, and sacral vertebrae respectively.

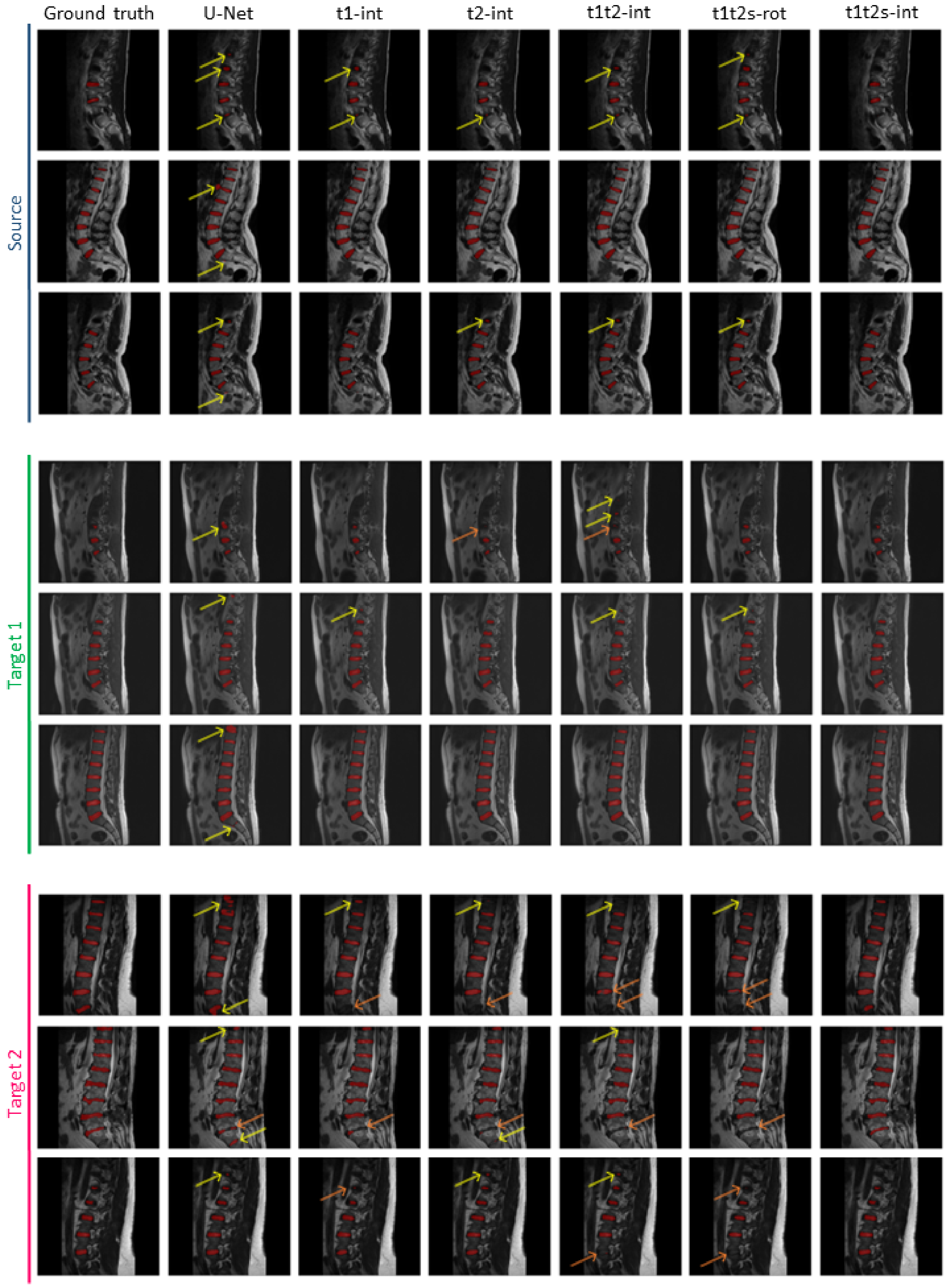


Figure 3.9: Qualitative results from the comparison of all the tested configurations on three random test images from each of the three datasets (source, target 1, and target 2 from top to bottom). In the source domain, the improvements brought by the proposed model (*t1t2s-int*) are particularly evident in the accurate segmentation of the contour of the discs. In target 1 and target 2, the proposed model demonstrates fewer false negatives, successfully segmenting all the discs present in the images. Yellow arrows in the image point to false positive predictions, while orange arrows to false negatives.

Table 3.2: Results of the performance metrics computed on the test sets of the three datasets, obtained from the baseline model (i.e. U-Net trained only on the Source dataset), the proposed model (*t1t2s-int*) and all the other dual-task models trained with various pretext task configurations. Values are reported as mean ± standard deviation.

Test dataset	Model	DSC (%)	HD (pixels)	Sen (%)	Spec (%)
S	U-Net	0.86±0.02	28.16±27.69	0.90±0.04	0.99±0.00
	t1-int	0.86±0.02	22.32±18.36	0.90±0.04	0.99±0.00
	t2-int	0.86±0.02	46.58±84.15	0.90±0.04	0.99±0.00
	t1t2-int	0.86±0.02	21.26±18.36	0.90±0.04	0.99±0.00
	t1t2s-rot	0.86±0.02	**21.25±18.36**	0.90±0.04	0.99±0.00
	t1t2s-int	0.86±0.02	21.69±23.68	0.90±0.05	0.99±0.00
$T1$	U-Net	0.89±0.03	33.47±37.15	0.88±0.08	0.99±0.00
	t1-int	0.91±0.03	24.74±40.92	0.92±0.07	0.99±0.00
	t2-int	0.91±0.04	32.82±40.23	0.90±0.07	0.99±0.00
	t1t2-int	0.90±0.05	33.20±39.20	0.90±0.08	0.99±0.00
	t1t2s-rot	0.88±0.06	27.33±41.72	0.83±0.10	0.99±0.00
	t1t2s-int	**0.92±0.04**	**13.59±17.70**	**0.90±0.07**	0.99±0.00
$T2$	U-Net	0.74±0.18	93.59±91.62	0.66±0.22	0.99±0.00
	t1-int	0.75±0.18	65.69±70.11	0.67±0.22	0.99±0.00
	t2-int	0.76±0.12	85.78±61.42	0.66±0.17	0.99±0.00
	t1t2-int	0.73±0.18	56.58±32.11	0.64±0.21	0.99±0.00
	t1t2s-rot	0.71±0.10	74.35±57.84	0.58±0.13	0.99±0.00
	t1t2s-int	**0.77±0.18**	**42.83±28.11**	**0.67±0.22**	0.99±0.00

Comparing the proposed DA model (*t1t2s-int*) with the baseline (U-Net), it becomes evident that the former consistently outperforms the latter on all the datasets and across each metric. For the source dataset S, the only one on which U-Net was trained, the performance in terms of DSC, Sen, and Spec are comparable for the two models. However, *t1t2s-int* achieves better results when considering the HD with an average value of 21.69 ± 23.68 pixels, outperforming U-Net in which HD has a mean value of 28.16 ± 27.69 pixels. For $T1$, the proposed DA model achieves a DSC of 0.92 ± 0.04 and an HD of 13.59 ± 17.70 pixels, outperforming the U-Net model, which achieves a DSC of 0.89 ± 0.03 and an HD of 33.47 ± 37.15 pixels. On the other hand, the proposed model shows comparable performance in terms of Sen and Spec, with values of 0.90 ± 0.07 and 0.99 ± 0.00, respectively, which are similar to those of U-Net. In the case of $T2$, the proposed DA model achieves a higher DSC of 0.77 ± 0.18 and a lower HD of 42.83 ± 28.11 pixels compared to the U-Net model, which has a DSC of 0.74 ± 0.18 and an HD of 93.59 ± 91.62 pixels. Similarly, the proposed model demonstrates a slightly higher Sens (0.67 ± 0.22) and Spec (0.99 ± 0.00) compared to U-Net. Fig. 3.6 vividly illustrates the qualitative results of this experiment, showing that the U-Net approach tends to produce sub-optimal disc segmentation, especially for discs situated in the outermost regions of the image, highlighting the superiority of the proposed DA model in these challenging regions. The comparison between U-Net and the proposed *t1t2s-int* is further explored in Fig. 3.7 and Fig. 3.8, which illustrate the results obtained by both models considering individual IVD segmentation. In Fig. 3.7 results for dataset $T1$ are reported, in which *t1t2s-int* shows overall better performances with higher median values and lower interquartile ranges (IQRs) for DSC, Spec and Sen. As regards HD, U-Net shows higher median values for the discs T10/T11, T12/L1, and L2/L3, with a wider IQR with respect to *t1t2s-int*. Fig. 3.8 displays the results for dataset $T2$. The performances are particularly lower with respect to $T1$, but with similar trends: *t1t2s-int* shows better performances than U-Net in terms of median values and IQRs for DSC, HD, and Sen. Spec, on the other hand, shows comparable median values for the two models for all discs, except for L2/L3, L3/L4, and L5/S.

To obtain the best setup of the pretext task, different training data configurations were evaluated. Results of the performance metrics, obtained using the different pretext tasks show that for the S dataset, all the tested models exhibited comparable performances. The *t1-int*, *t2-int*, and *t1t2-int* models achieved mean DSC values of 0.86 ± 0.02. The proposed model demonstrated consistency in performance across the different pretext tasks. All training strategies performed quite well on $T1$, achieving notable peaks of 0.92 ± 0.04, 13.59 ± 17.70, 0.90 ± 0.07, and 0.99 ± 0.00 for DSC, HD, Sen, and Spec, respectively, when using the *t1t2s-int* configuration. The *t1-int* model achieved a mean DSC of 0.91 ± 0.03, a mean HD of 24.74 ± 40.92 pixels, a mean Sen of 0.92 ± 0.07, and a mean Spec of 0.99 ± 0.00. The *t2-int* and *t1t2-int* models demonstrated slightly improved results in terms of DSC, with values of 0.91 ± 0.04 and 0.92 ± 0.04, respectively, however the HD results were higher for both these models if compared with *t1-int*. Similar trends are observed for $T2$, even though with lower mean values of the metrics compared to $T1$. Also in this case the *t1t2s-int* model achieves the highest DSC (0.77 ± 0.18) and the lowest HD (42.83 ± 28.11 mm) compared to the other tested models. The *t1-int* model achieved a mean DSC of 0.74 ± 0.20, while the *t2-int* and *t1t2-int* models obtained slightly higher mean DSC values of 0.76 ± 0.12 and 0.73 ± 0.18, respectively.

In the experiment comparing the performance of the intensity prediction pretext task (*t1t2s-int*) and the rotation angle prediction pretext task (*t1t2s-rot*), slight differences were observed among the three datasets regarding its effectiveness within the dual-task approach. For the S dataset, the *t1t2s-rot* configuration showed a similar mean DSC of 0.86 ± 0.02 and HD of 21.25 ± 18.36 with respect to *t1t2s-int*. For $T1$ the *t1t2s-rot* configuration obtained the lowest performances in terms of DSC and Sen, with mean values of 0.88 ± 0.06 and 0.83 ± 0.10, respectively. In the dataset $T2$, introducing the rotation as pretext task results in a deterioration of performance if compared to the baseline (U-Net), with DSC = 0.71 ± 0.10, HD = 74.35 ± 57.84, and Sen = 0.58 ± 0.13. Qualitative results shown in Fig. 3.9 further support the effectiveness of the proposed model in accurately segmenting the IVD while minimizing the presence of segmented spots outside the designated region.

3.4 Discussion

IVD localization and segmentation have become the focus of extensive research in the medical image analysis community due to their significance in identifying various spine-related pathologies [221]. While DL algorithms have been employed to address IVD segmentation, these models frequently encounter performance constraints when confronted with data from diverse domains, including specific acquisition devices and protocols or patients with varying medical conditions [222]. The difficulties in segmenting IVD in MRI stem from the wide range of variations in their appearance, including signal intensity, shape, and texture. Furthermore, the challenge of accurately segmenting the discs is exacerbated by two factors: the limited number of voxels allocated for their representation reduces the discriminative information available, while the partial volume effect further complicates this task by causing the disc boundaries to blend with neighboring tissues. The main aim of this research is to examine the possibility of using self-supervised learning to perform unsupervised DA for IVD segmentation. Specifically, the aim is to investigate whether the use of pretext tasks can assist in this process. The reason behind this approach is the belief that the model will learn representations during pretext task training that are not domain-specific and can be applied to the target domain. Specifically, a novel approach was adopted by designing a pretext task focused on predicting specific intensity variations, such as Gaussian noise, Gaussian blur, and contrast enhancement. This choice was made instead of more traditional methods like rotation degree prediction, primarily because of its relevance to the unique challenges of IVD segmentation. Intensity variations commonly occur in medical images acquired from different sources and can significantly hamper the performance of segmentation models by introducing variations in contrast and brightness. In light of the challenges posed by intensity variations, the analysis of intensity histograms across the three datasets serves as a pivotal insight. As observed in the intensity histograms reported in Fig. 3.3, the distribution of intensities in the S dataset differs markedly from that of the target datasets $T1$ and $T2$. This pronounced peak in the lower intensity range of the S dataset could be indicative of unique acquisition conditions, peculiarities of the MR scanner, or particular patient populations. On the other hand, the more nuanced and diverse intensity profiles of the target datasets highlight that aligning the domains is not a straightforward task. This analysis, grounded in the intensity histograms, underscores the importance of developing DA methods that address not just the structural differences, but also the nuanced variations in intensity characteristics inherent to MR datasets. The proposed pretext task, tailored to address these issues, enables the model to effectively learn intensity-invariant features. As a result, the model achieves robust segmentation performance across a source and two target datasets containing patients with different pathological conditions and acquired with various scanning devices (namely Philips Ingenia, Siemens, Philips Achiva, Philips Elition GE Sigma).

Evaluating the proposed strategy against the baseline model (U-Net), *t1t2s-int* reached the best performance metrics in both $T1$ and $T2$ while maintaining the same performances with respect to U-Net in S, as shown in Table 3.2. This indicates that the strategy not only excelled in the intended domains but also managed to preserve the effectiveness it had demonstrated in the source domain. This aspect is crucial as it ensures that implementing the strategy does not result in any detrimental effects or a decrease in performance in the original domain. By examining the performance across individual discs, Fig. 3.7 clearly demonstrates how the proposed methodology outperforms the others in terms of stability. Specifically, it exhibits consistently favorable results in median and IQR across different IVD on $T1$. Notably, even for the segmentation of challenging IVD such as T10/T11 and L5/S, which are typically more difficult due to their position in the image, the proposed methodology maintains its superior performance. This highlights the robustness of the proposed approach, making it particularly well-suited for accurate and reliable IVD segmentation across different locations within the image. In the case of $T2$ the scenario is slightly different, as the complexity that characterizes this dataset is reflected in the variability of the results between one disc and another. In particular, the discs T10/T11, T12/L1, and L5/S resulted in the lowest DSC for both U-Net and *t1t2s-int*, even though this latter outperforms the other in all cases. A similar trend can be observed for HD and Sen.

The best training strategy is achieved when incorporating intensity pretext tasks across multiple datasets, including the two target domains and the source domain. This is confirmed by comparing the proposed model with other pretext configurations. In fact, it reached DSC = 0.92 ± 0.04, HD = 13.59 ± 17.70, Sen = 0.90 ± 0.07 for $T1$ and a DSC = 0.77 ± 0.18, HD = 37.63 ± 25.40, Sen = 0.90 ± 0.07 for $T2$. When dealing with pretext applied on only a specific target dataset (*t1-int* and *t2-int*), the performances are better for the dataset on which the pretext was carried out with respect to the other datasets on which the pretext task is not applied, as can be seen in Table 3.2. Furthermore, it is important to note that Spec achieved consistently high values of 0.99 in all experiments, underscoring the model's ability to correctly identify negatives across all cases. A similar trend is observed when applying the pretext task to the two target datasets and not to the source dataset (*t1t2-int*). This behavior is expected as the model has learned features from just one specific dataset while *t1t2s-int* improves the generalization performance and adaptability across diverse data distributions. This is also evident when qualitatively evaluating the results, as can be observed from Fig. 3.9. The reduced presence of segmented spots outside the true label region is a critical advancement in IVD segmentation, as it improves the reliability of the segmentation results. This outcome is particularly important in medical applications where precise delineation of the anatomical region of interest (IVD in this case) is crucial for accurate diagnosis and treatment planning.

Comparing the proposed intensity pretext task (*t1t2s-int*) with a more traditional one (*t1t2s-rot*), which involves the prediction of fixed rotation angles, the superiority of the former was demonstrated in the context of self-supervised DA for medical image analysis. The problem under investigation, in fact, may have inherent characteristics that make intensity information more important for the target task. In fact, intensities can play a crucial role in distinguishing between IVD and vertebral bodies in images acquired from different devices. Moreover, it is worth highlighting that different MRI scanner manufacturers often have proprietary calibration methods, sequences, and post-processing software. These variations can introduce differences in intensity distribution profiles and contrast settings across devices. This variability across manufacturers further emphasizes the importance of accounting for these differences when analyzing images from multiple sources.

While the proposed model exhibited promising results, it is important to recognize certain limitations. Firstly, the evaluation was conducted on a testing dataset with a limited number of samples. Employing larger datasets could enhance the evaluation of statistical significance and enhance the model's ability to generalize to diverse populations and imaging protocols. Additionally, the primary objective was to examine the feasibility of the approach; therefore, only a standard U-Net was employed for evaluation. Further advancements could be made in this regard, such as utilizing pyramid structures to construct multi-scale representations and address multi-scale variations or leveraging self-attention modules [223] to improve performance. Future developments include firstly the use of a wider variety of transformations in the pretext task, which could be investigated to enable the model to learn more comprehensive and adaptable features, improving its ability to handle different types of image data and enhancing segmentation performance. In fact, it has been demonstrated [213] that pretext tasks such as predicting intensity histograms, shapeless local patch discrimination and texture discrimination can enable the model to learn intensity and texture-aware features. Further advancement can be obtained through the introduction of an additional adversarial loss [224]. Adversarial training has proven, in fact, to be effective in reducing domain shift and improving model generalization. By incorporating an adversarial loss into the training process, the model could potentially become more robust to domain differences, leading to enhanced segmentation performance across various datasets.

3.5 Conclusion

This study proposed the first multi-target unsupervised DA approach for IVD segmentation, consisting of a dual-task segmentation model based on intensity distortion pretext tasks. The proposed model, which simultaneously segments the IVD region and predicts intensity transformations, was trained on unlabeled data from multiple domains, acquiring domain-invariant features and demonstrating robust performance across diverse datasets, thus overcoming challenges associated with intensity variations arising from different acquisition devices and medical conditions. By leveraging available unlabeled data, the model achieves superior segmentation performance with respect to supervised segmentation models (U-Net) which often rely on large annotated datasets. The successful application of unsupervised DA in the context of medical image analysis is a significant contribution of this study, and the proposed approach holds the potential to be adapted to address other segmentation challenges, particularly those instances where DL models falter due to domain discrepancies among images acquired from different devices.

Chapter 4

Artificial intelligence-driven applications in ultrasound images archives

US imaging, also known as sonography, is a non-invasive medical imaging technique that employs high-frequency sound waves to create real-time images of the internal structures of the human body [225]. US image acquisition involves the transmission of sound waves into the body using a transducer. The waves, upon encountering different tissues and structures, are reflected back to the transducer at varying frequencies and intensities. These echoes are then converted into electrical signals, which are processed to form the images visualized on the US device [226]. US imaging is valued for its safety, absence of ionizing radiation, cost-effectiveness, and real-time imaging capabilities [225]. It is commonly used in various medical domains, including obstetrics and gynecology, cardiology, and diagnostics of abdominal and musculoskeletal conditions.

However, despite the well-recognized clinical utility, US imaging presents unique challenges such as the high dependence on the sonographer's diagnostic experience, and high inter- and intra-observer variability across different institutes and US systems manufacturers. Moreover, US images can be subjected to low imaging quality, intensity inhomogeneities, presence of shadows, and high noise levels which hinder image interpretation.

To overcome these challenges, different automated image analysis methods have been proposed over the years with the aim of making US diagnosis and assessment more objective and accurate. In this landscape, DL has found applications in automating the analysis of US images, as a support in the diagnosis, solving a variety of tasks like biometric measurements, image segmentation, tissue characterization, and disease detection.

Among the aforementioned tasks, this Chapter will outline a segmentation approach. US image segmentation is often used to quantitatively analyze clinical parameters such as the volume and shape of organs, detect and classify lesions (e.g., breast, prostate, thyroid nodules), and extrapolate biometric measurements (ranging from infant to elderly).

Recently, US saw a rise in usage for musculoskeletal assessments after undergoing considerable technological enhancements, as AI-based computer-aided detection and computer-aided diagnosis are fundamental to ensure less costly, more effective and improved US imaging [227]. In this field, the most investigated DL tasks are nerve identification and segmentation [228, 229], myositis classification [230] and synovitis classification [231], nevertheless, the field of musculoskeletal applications still represents a largely unexplored domain in numerous aspects.

Within this scenario, this Chapter describes a new method to automatically localize and segment the median nerve at the inlet of the proximal carpal tunnel, to measure the nerve Cross-sectional Area (CSA) in US images.

US imaging is recognized as a useful support for Carpal Tunnel Syndrome (CTS) assessment through the evaluation of median nerve morphology. However, US is still far to be systematically adopted to evaluate this common entrapment neuropathy due to US intrinsic challenges, such as its operator dependency and the lack of standard protocols. Relying on archived US images acquired from 103 rheumatic patients, this Chapter addresses the problem of automatic median nerve segmentation and CSA measurement by designing a fully-automatic DL approach. The framework consists of a Mask R-CNN with two additional transposed layers at the segmentation head to accurately segment the median nerve directly on transverse US images, and calculate the CSA of the predicted median nerve.

This work demonstrated its ability to address the problem of median nerve segmentation in US rheumatological images with DL, paving the way for future research in the field. From a clinical perspective, the proposed framework proved to be a valuable tool to support the clinical routine increasing the reproducibility of CSA measurements.

A journal paper on this work has been published as [232]: Di Cosmo, M., Fiorentino, M. C., Villani, F. P., Frontoni, E., Smerilli, G., Filippucci, E., and Moccia, S. (2022). A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet. Medical & Biological Engineering & Computing, 60(11), 3255-3264.

4.1 Introduction to a deep learning approach for median nerve evaluation in ultrasound images of carpal tunnel inlet

CTS accounts for 90% of peripheral entrapment neuropathies, affecting up to 5% of the general population [233]. This condition occurs when the median nerve is compressed at the wrist as it passes through a narrow osteofibrous canal along with the nine finger flexor tendons [234]. The median nerve stretches, compresses and translates in response to upper extremity motion, but in patients with CTS its mobility is restricted, which indicates nerve dysfunction [234].

Traditionally, the diagnosis of CTS relies on clinical history and physical examination [235], sometimes investigated further with electrodiagnostic tests, sensitive in examining nerve conduction and eventual damages [234]. Aside from electrodiagnosis, which is expensive, time-consuming and presents limited ability to predict CTS severity or intervention outcome [236], US imaging can also be used. In assessing CTS, US allows to detect structural anomalies through the direct visualization of the nerve, its position and morphology: in fact, altered shape of the median nerve due to the compression of the surrounding nonrigid structures is expected in CTS patients [233].

Among the US parameters which can be evaluated from the carpal tunnel, the most common and reliable is the CSA of the median nerve measured at the proximal carpal tunnel. However, the CSA measurements are currently performed on US relying on a hand tracing method, and their cut-off values for CTS diagnosis vary widely, ranging from 9 to 14 mm^2 [233].

US imaging presents unique challenges to be faced: it is highly dependent on sonographer's experience, and subjected to high inter- and intra-observer variability across different manufacturers' US systems. Moreover, US images can be subjected to low imaging quality, intensity inhomogeneities,

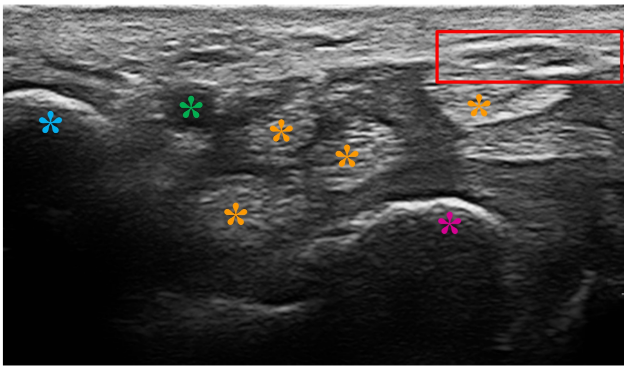


Figure 4.1: Ultrasound transverse scan sample acquired at the proximal carpal tunnel inlet. A red box includes the median nerve section; asterisks of different colors mark other relevant structures: pisiform bone profile in blue, semilunar bone profile in purple, ulnar artery in green, digital flexor tendons in yellow.

presence of shadows and high noise level. In addition, in carpal tunnel imaging the median nerve identification is made harder by the presence of many rounded structures, such as the wrist bones, transverse carpal ligament and digital flexor tendons, and by nerve morphological variations in relation with disease severity, with other concomitant pathologies and also with height, sex, weight and age of the subjects [233]. A sample of carpal tunnel US image is shown in Fig. 4.1.

To address these challenges, the development of advanced automatic US image analysis methods is essential to make US a more objective and accurate support tool for CTS assessment. In this respect, DL has already shown its huge potential for medical US analysis [237]. At present, multiple types of deep networks, especially CNN, have been successfully involved in various US images tasks, such as lesion and nodule classification, object detection and anatomical structure segmentation [237], thus implying DL potentiality to improve and standardize even CTS diagnosis through an automatic median nerve section identification.

Embracing this idea that DL may provide reliable support to sonographers, this study extends a previous preliminary work in [229] proposing the following contributions:

1. Development of an end-to-end CNN, i.e. a Mask R-CNN [166], for localization and segmentation of the median nerve at the inlet of the proximal carpal tunnel, further improved by the insertion of two additional transposed layers at segmentation head.
2. A comprehensive study conducted on transverse US images acquired in daily clinical practice.
3. Evaluation of CSA measurement based on the median nerve section segmented by the algorithm in comparison with manual tracing of nerve boundary performed by expert sonographers.

4.1.1 Related work

Several studies faced the median nerve segmentation problem from US imaging involving model-based approaches. In [238], the phase-based probabilistic gradient vector flow algorithm was used to track sciatic nerve region, obtaining an average DSC of 0.90. The work in [239], instead, proposed the adaptive median binary pattern as the texture feature of a tracking algorithm with an Acc of 95%. A segmentation pipeline including a pre-processing stage (filtering, de-noising, contrast enhancement), features extraction in a Region of Interest (ROI) and a SVM classifier was proposed in [240]. This method generated an average DSC of 0.81. However, even with good results, these approaches are parameter sensitive and require a certain degree of time-consuming manual intervention, especially for selecting the initial contour, thus possibly leading to segmentation errors.

After DL has emerged as leading ML tool in various research fields, including medical US analysis, recently some researches approached the median nerve segmentation involving CNNs implementation. [241] combined a CNN, which detects the ROI around the nerve, with the probabilistic gradient vector flow method to delineate the median nerve contour on a dataset composed of US images extracted from 10 videos, each with 500 frames, from 10 patients. The results revealed an average DSC of 0.85. In [242] the U-Net architecture [72] was used to identify the median nerve in the brachial plexus in US images, which were all pre-processed using linear Gabor binary patterns before being supplied to the U-Net for segmentation. They obtained an average DSC of 0.67, thus considering that the use of U-Net to directly segment the median nerve is not effective. In [236], a multi-input similarity CNN was proposed to track the median nerve in US videos from 50 patients, who where asked to perform specific wrist motions. 100 US videos of 6 seconds, each with 180 frames, were involved in this study, in which one target ROI containing the median nerve, manually defined in the first frame, is compared with candidate search images to find the more similar on the next frame of image stack. It is worth noticing that this method relies on the manual identification of ROIs from expert clinicians as input to the model, which is a relevant limitation. [243] proposed a fully DL framework based on U-Net for the localization and segmentation of the median nerve in US image sequences. The model, called DeepNerve, integrates also a MaskTrack [244], a video object segmentation technique, and a convolutional long short-term memory [245], to process temporal information. Six patients were involved and a total of 24 videos, each with 420 frames and lasting 17.5 seconds. The images of the US sequences were cropped around the median nerve before being used to train and test the model. DeepNerve overcame the segmentation performances of the conventional active contour model, generating an average DSC value of 0.89. However, this method used images cropped around the median nerve as input, and the small number of patients involved limited the anatomical variability considered in the study. Even in a recent work by [246], two implementations of the U-Net model were considered on a dataset of 505 videos with 5560 annotated frames acquired involving 99 patients (with an average of 5,1 videos): one model was based on single-frame segmentation, the other was made using focus windows and spatial information from the previous segmented frame to redirect the focus of the search area for the next frame. The best results were achieved by the latter model with an average DSC of 0.88, but requiring the first frame manual definition by a user and ROIs as input to the model. Despite the promising results, the main limitation of these DL methods is that they require the manual identification of a ROI around the median nerve, and this poses issues relevant to time consumption and inter-clinician variability. The work conducted by [247] evaluated the performance in median nerve segmentation of different DL models, including DeepLabV3+, U-Net, Feature Pyramid Network (FPN) and Mask R-CNN [166], on US image sequences acquired from 36 subjects. The best performances were achieved by the Mask R-CNN with an IoU score close to 0.83. This work, however, focused on a small variety of anatomy and excludes unusual morphologies.

In accordance with results achieved by [247], and in contrast with the other DL approaches found in the literature on this field, in which U-Net-based models were chosen to face this task, in this work the median nerve segmentation was approached by implementing a Mask R-CNN,

	Acquisition site	Frame sequences	N. of US images	N. of patients
[242]	Brachial plexus forearm	No	11508	
[236]	Carpal tunnel	Yes (100)	18000	50
[243]	Carpal tunnel	Yes (24)	10080	6
[246]	Proximal carpal tunnel inlet	Yes (505)	5560	99
[247]	Proximal carpal tunnel inlet	Yes (36)	18625	36
Proposed Model	Proximal carpal tunnel inlet	No	246	103

Table 4.1: Overview of the Ultrasound (US) dataset characteristics in Deep Learning (DL) literature for median nerve segmentation, in terms of US acquisition site, dataset size (frames selection or frame sequences, total number of images) and patients involved in the study.

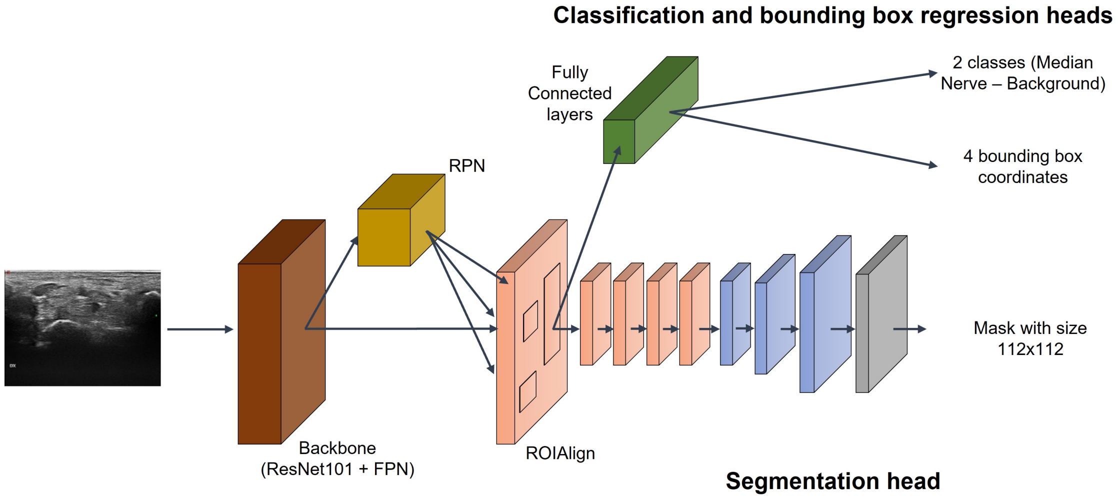

Classification and bounding box regression heads
Fully Connected layers
2 classes (Median Nerve – Background)
4 bounding box coordinates
RPN
Mask with size 112x112
Backbone (ResNet101 + FPN)
ROIAlign
Segmentation head

Figure 4.2: Schematic representation of model architecture, composed by a backbone, Region Proposal Network (RPN), and the three heads for classification, bounding-box regression and segmentation, all fed from the ROIAlign with 100 ROI candidates. The segmentation head is represented more in details as it was provided with two additional transposed layers compared with original Mask-RCNN.

which simultaneously detects target objects in the image and from that generates a high-quality segmentation mask for each instance. The aim is to provide a unified framework, which does not involve preliminary ROI identification or parameter-sensitive procedures.

In addition, the dataset used for the present study is significantly different from the ones described in the state-of-art [236, 242, 243, 246, 247]: the focus here is on the morphology rather than the motion of the median nerve, thus considering US single frames instead of full-frame sequences and involving in the study a greater number of patients, covering a higher anatomical variability. Table 4.1 summarizes the characteristics of these data sets and highlights the differences with the dataset used for this analysis.

The following subsections present and discuss the proposed approach in detail.

4.2 Materials and methods

In this study, the median nerve segmentation from transverse US images acquired at the proximal carpal tunnel inlet is approached by deploying an end-to-end DL algorithm based on a Mask R-CNN implementation [166]. A schematic representation of the model proposed is shown in Fig. 4.2.

ResNet-101 [59] was used as backbone in combination with the FPN [167], allowing median nerve detection at multiple scales, which improves the performance of semantic segmentation over relying on a single scale analysis. As in the original implementation by [166], the RPN is used to generate proposals, i.e. rectangular regions in the US image with a high probability of containing the median nerve, which are predicted starting from anchors, which are here built with 5 different sizes and 3 different scales. The selected proposals are processed by the ROIAlign layer, which resizes the proposals to a constant $d \times d$ output matrix before feeding them to the heads.

The classification and regression heads are both made of two fully-connected layers with 1024 neurons and an additional third fully-connected layer, which has 2 neurons followed by a softmax function to predict the proposal class (i.e., median nerve or background) for the classification head and 4 neurons, linearly activated, to predicted the anchor box correction factors in the regression head. The segmentation head, instead, consists of four 3x3 convolutional layers with 256 filters, each activated with the ReLU, and three transposed convolutions with 256 2x2 filters, ReLU activated, which allow to recover spatial resolution up to 112x112. In this work, architectural changes from the original Mask R-CNN are introduced at the segmentation head to improve output mask resolution. In fact, the use of three transposed convolution layers instead of only one, as in the original Mask R-CNN, allows to increase the output resolution and to deal with the fragmented and low-contrasted edges of the median nerve. To obtain stable convergence, the last layer performs a 1x1 convolution and it is activated by a Sigmoid function.

The proposed method was trained and tested using TensorFlow on a GPU GeForce RTX 2080.

4.2.1 Dataset

For this study, 103 patients with rheumatic and musculoskeletal disorders were recruited at the Rheumatology Unit of "Carlo Urbani" Hospital in Jesi (Ancona, Italy). All patients signed informed consent and the data acquisition was conducted in compliance with the Helsinki Declaration and with the approval of the local ethics committee (Comitato Etico Regione Marche, number 262). The US assessment was carried out using a MyLab Class C (Esaote SpA, Genoa, Italy) US system equipped with a 6–18 MHz linear probe taking transverse scans in accordance with the 2017 EULAR standardized procedures for US imaging in rheumatology [248]. US images at the proximal carpal tunnel inlet were acquired bilaterally from the patient wrists with the forearm resting supine on the examination bed and fingers in a neutral position. The number of images per patient is variable, but of the same order of magnitude, and the few cases in which more than one image is acquired from the same patient were carefully considered as part of the same set (either training or testing). 22 out of 103 patients (21%) had a clinical diagnosis of CTS and some anatomical variants were observed and included in the study. The presence of the following anatomical variants was registered: bifid median nerve, persistent median artery, and accessory muscles within the carpal tunnel. The images composing the dataset were acquired by three sonographers with different degrees of experience in the musculoskeletal US (G.Sa.: 1 month with dedicated intensive training; G.Sm.: 4 years; E.Fi.: more than 20 years of experience). Images considered of insufficient quality were excluded from the dataset after a revision made by the expert sonographers. Manual annotation was performed by one sonographer (G.Sa.) under the supervision of the other two.

The annotations were used as ground truth for the training of the CNN proposed for the segmentation task. The dataset included a total of 246 US images with size equal to 606x468 pixels. The images with the corresponding masks were resized to 512x512 pixels using bilinear interpolation. In addition, the images were zero-padded at the right-most and bottom-most edges to get squared images with a size multiple of 32, as required by the FPN, while keeping the aspect ratio unchanged.

4.2.2 Experimental setup

The dataset was randomly split by patients, whose demographic and clinical characteristics matched inclusion criteria designed by rheumatologists before performing the acquisition. To cope with the small amount of data available, 5-fold cross-validation was performed. All the ablation studies and the comparison with the state-of-the-art models were conducted training in 5-fold cross-validation and testing on the model with the best validation loss.

Considering the relatively small size of the dataset and to reduce the chances of overfitting, during training data augmentation was performed on the fly by randomly scaling 80% to 120% of the original size, and translating -20% to 20% on both x- and y-axis independently, and performing random rotation between (-10°, 10°) and shearing between (-2°, 2°). The ranges for the affine transformations were chosen to ensure that the nerve remains always visible in the images.

To improve training speed and accuracy, transfer learning was performed initializing all the layers of the model except for the input layers of the network heads with weights computed on the COCO dataset [168]. Freezing the backbone while focusing on the training of the network heads aimed to increase the feature extraction process through the support of a large natural image dataset.

The training was performed following guidelines for training CNNs, including dropout and weight decay as regularizers. Stochastic Gradient Descent was deployed as optimizer for 150 epochs with a learning rate of 0.001 and momentum of 0.9. A total of 256 anchors per image were used, with varying sizes (32, 64, 128, 256, and 512) and aspect ratios (1:1, 2:1, 1:2). These values were chosen considering the median nerve section dimension. The ROIAlign resized proposals to a fixed output size of 14x14 considering a total of 100 training ROIs per image, as a trade-off between accuracy and memory consumption.

The network was trained defining a multi-task cross-entropy loss on each ROI combining the loss of classification, localization, and segmentation mask equally weighted: $L = \alpha L_{cls} + \beta L_{bbox} + \gamma L_{mask}$, where L_{cls} and L_{bbox} are class and bounding box losses of Faster R-CNN, respectively, and L_{mask} is the mask loss defined in [249], and α, β and γ are constants, which we set to 1 after experimental investigations.

In addition, from the obtained median nerve segmentation, the CSA was calculated knowing that a single pixel in the US images of the dataset has a dimension equal to $0.062mm \times 0.062mm$. The CSA was calculated only on TP predictions.

4.2.3 Comparison with literature and ablation studies

As mentioned in Subsection 4.1.1, a relatively small number of studies is focused on DL application on US for CTS assessment and in most of these contributes, as in [242], [243] and [246], U-Net models were chosen to get the median nerve segmentation. Hence, even though the dataset used for this study is composed of still US images instead of US videos as in current literature, and thus these works are not directly superimposable, a performance comparison was performed among the proposed model and some U-Net based approaches.

This comparison aims to prove the effectiveness of the deployment of a Mask R-CNN architecture rather than U-Net models to obtain an end-to-end framework, which accurately segments the median nerve without the requirement of any a priori localization or parameter-sensitive post-processing.

The architectures chosen for the comparison were the one deployed in [243] of the U-Net, which kept the original implementation on this state-of-art network from [72], and a Lightweight U-Net, in which the network's depth was reduced from 5 to 4 layers and batch normalization was used as a follow-up step to the first convolution in each layer to avoid premature convergence. To evaluate the best performances of these models in comparison with the proposed one, the train was performed using the Binary Cross-Entropy (BCE) loss, which is the default loss for segmentation models, and also combining the BCE loss with the DSC loss (BCE-DSC loss), expected to provide more stability to the models [250]. The DSC is also the metric mainly used to evaluate the model performance in terms of segmentation, which was calculated in this work as previously defined in Eq. 2.3.

In the ablation study, it was investigated whether a larger backbone network results in increased accuracy: thus, the Resnet-101 combined with FPN (ResNet-101-FPN) was compared with the ResNet-50 FPN. To evaluate if the augmentations applied lead to a greater generalization of the model, the ablation studies also included an experiment training the model without any type of augmentation. In addition, it was evaluated the effect of having a different number of transposed convolutions in the segmentation head. This was done to assess the effects of an increased resolution of the output of the segmentation head on the overall segmentation performance. The segmentation head was tested with one (Mask28) and two (Mask56) transposed convolutional layers, leading to the output size of the head of 28x28 and 56x56, respectively. For a fair comparison, the ablation studies were performed using 5-fold cross-validation, the same training settings, and computational hardware.

4.2.4 Performance metrics

Prec, Rec, and mAP are used to evaluate the performance in median nerve localization. Prec and Rec were computed as indicated by Eq. 2.7 and Eq. 2.5, respectively. A prediction was considered as a True Positive (TP) if the detected bounding box overlapped the bounding box surrounding the ground truth segmentation for at least 70% and had confidence higher than 0.98. A wrong positive detection was considered as a False Positive (FP), in which the predicted bounding box did not reach 70% of the overlapping threshold with the ground truth bounding box. A False Negative (FN) was considered when the actual instance was not detected, thus no bounding box was predicted at all. The value of 70% as the threshold for defining TP, FP, and FN was chosen to provide more strict and reliable segmentation from the nerve detection: the standard Pascal VOC evaluation practice [251] with minimum overlapping at 50% between predicted and ground truth bounding boxes was considered as not accurate enough for properly measuring CSA, a fundamental parameter for CTS diagnosis. mAP, which represents the average of the area under the Prec-Rec curve, was also computed. The median nerve segmentation performance was measured using the DSC, as defined in Eq. 2.3.

In addition, the CSA was automatically calculated from the median nerve section predicted by the algorithm, knowing the dimensions of a single pixel ($0.062mm \times 0.062mm$) in the US images. The CSA was calculated only on TP predictions and compared with manual measurements performed

Table 4.2: Performance evaluation metrics in terms of mean value and standard deviation. Mean Average Precision (mAP), Recall (Rec), Precision (Prec), and Dice Similarity Coefficient (DSC) are reported for the proposed model and the ablation studies conducted over it: Mask-R50 is the model trained using as backbone Resnet50 combined with FPN; NoAug is the model trained using no augmentations on the training data; Mask28 and Mask56 are variants of the model with a different output resolution from the segmentation head, including one and two transposed convolutional layers, respectively.

	mAP	Rec	Prec	DSC
Mask-R50	0.89 ± 0.27	0.88 ± 0.27	0.86 ± 0.26	0.84 ± 0.21
NoAug	0.89 ± 0.24	0.90 ± 0.29	0.87 ± 0.31	0.84 ± 0.25
Mask28	0.91 ± 0.36	0.92 ± 0.25	0.87 ± 0.28	0.82 ± 0.26
Mask56	0.93 ± 0.23	0.89 ± 0.28	0.89 ± 0.27	0.84 ± 0.22
Proposed Model	0.94 ± 0.23	0.94 ± 0.23	0.92 ± 0.24	0.87 ± 0.20

by the sonographers measuring the Mean Absolute Error (MAE).

4.2.5 Statistical analysis

The Kolmogorov-Smirnov test was performed to assess if the data were normally distributed, using an α value of 0.05. As the data are non-normally distributed (the p-value of Kolmogorov-Smirnov test is equal to $0.48 \times e^{-143}$), a Mann-Whitney test with $\alpha = 0.05$ was performed to compare the CSA measurements.

The agreement in the CSA measurements between the sonographer annotation (i.e., the gold standard) and the algorithm was calculated using a two-way mixed-effects Intra-class Correlation Coefficient (ICC) with 95% Confidence Interval (CI). The ICC is regarded as excellent if above 0.9, as good if between 0.75 and 0.9.

4.3 Results

The proposed model demonstrated effective performance in both detection and segmentation of the median nerve section, with average metrics results of mAP, Rec, Prec, and DSC equal to 0.94 ± 0.23, 0.94 ± 0.23, 0.92 ± 0.24 and 0.87 ± 0.20, respectively. The average inference time for each image on a GPU GeForce RTX 2080 was 1.7 s, which could be further improved with more powerful computational resources.

Table 4.2 summarizes the results obtained by modifying the model architecture using a different backbone (Mask-R50) and considering two different output resolutions of the segmentation head, leading to masks with sizes 28x28 (Mask28) and 56x56 (Mask56).

To evaluate the segmentation capability, the proposed model was compared with the U-Net and Lightweight U-Net models deployed in literature, referring in particular to [243]. Table 4.3 outlines the segmentation performances of these models in terms of DSC, expressed as mean ± standard deviation value. Visual samples are shown in Fig. 4.3: a sample of a healthy median nerve section (Fig. 4.3a), a sample acquired from a patient with CTS (Fig. 4.3b), a sample containing a prominent persistent median artery (Fig. 4.3c) and a sample of a bifid median nerve (Fig. 4.3d). Moreover, the CSA was measured on the predicted median nerve sections. Without considering FP and FN predictions, the values were comparable with the ones manually measured by the sonographer with a MAE of 0.92 mm^2. On average, CSA measured by the sonographer was 10.36 ± 4.52 mm^2, while

Table 4.3: Comparison of segmentation performance in terms of Dice Similarity Coefficient (DSC) of the proposed model and of the U-Net and Lightweight U-Net trained using two different losses, i.e. the Binary Cross Entropy (BCE) loss and the BCE-DSC loss.

	DSC
U-NET (BCE loss)	0.78 ± 0.23
U-NET (BCE-DSC loss)	0.82 ± 0.20
Lightweight U-NET (BCE loss)	0.78 ± 0.19
Lightweight U-NET (BCE-DSC loss)	0.76 ± 0.22
Proposed Model	0.87 ± 0.20

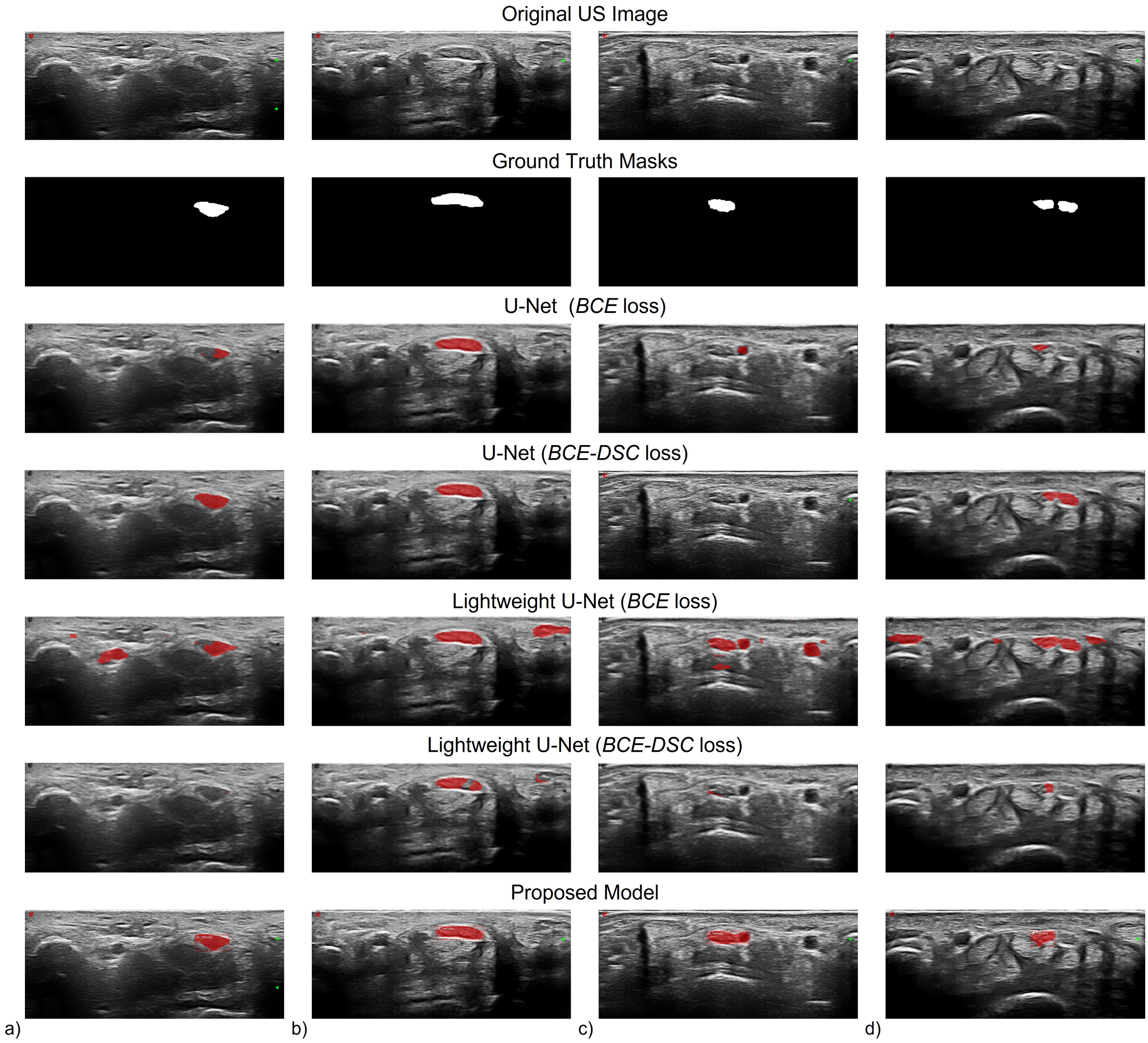


Figure 4.3: Four visual samples of the median nerve section. From top to bottom row: original Ultrasound (US) image, ground truth mask, U-Net trained with Binary Cross Entropy (BCE) loss prediction, U-Net trained with Binary Cross Entropy-Dice Similarity Coefficient (BCE-DSC) loss prediction, Lightweight U-Net trained with BCE loss prediction, Lightweight U-Net trained with BCE-DSC loss prediction, proposed model prediction. For displaying purposes, only the upper part of the US image, which contains the median nerve section, is shown.

CSA automatically calculated from the predicted segmentation masks was 10.38 ± 4.24 mm^2, with no significant difference ($p=0.88$). The agreement between the automatic algorithm measurement and the sonographer manual measurement of the CSA is remarkable [ICC 0.97 (95% CI 0.94–0.98)].

4.4 Discussion

Despite the increasing interest in US support for CTS assessment and the well-established usefulness as a confirmatory diagnostic test of the median nerve size measurement, US imaging is still struggling to be regularly employed in diagnostic work-up. This is partially due to the high competence required to perform and interpret US at the carpal tunnel level, the lack of protocol standardization, and the high variability among sonographers' evaluations. Therefore, this work proposes an end-to-end DL approach to support sonographers for median nerve compression evaluation. Specifically, the median nerve segmentation was approached by developing a Mask R-CNN model, obtaining remarkable results for both localization (mAP=0.94 ± 0.23, Rec=0.94 ± 0.23, Prec=0.92 ± 0.24) and segmentation (DSC=0.87 ± 0.20). Moreover, the automatic measurement of the CSA from the predicted median nerve section resulted to be in agreement with the manual measurement of the CSA (with an average MAE of 0.92 mm^2), implying the possibility of reducing reliance on the sonographer's expertise in carpal tunnel US evaluation while increasing intra- and inter-observer reliability.

Differently from other semantic segmentation models, Mask R-CNN solves the segmentation problem on top of localization, producing a mask for each recognized object, instead of just one final mask, thus leading to more accurate results. Previous works, in fact, approached the problem deploying U-Net based models [242, 243, 246], but they all involved some manual intervention in ROI identification or nerve contour definition to obtain good median nerve segmentation. The most similar work from a methodological point of view is the one from [247], in which the best results are achieved implementing a Mask R-CNN model; however, even with less data, the proposed model achieved higher performance on the collected dataset, which includes a greater number of patients and thus a higher variability, confirming the instance segmentation as more suitable and better performing than semantic segmentation approaches.

Therefore, the proposed model was compared with different implementations of U-Net models proving the better outcomes reached, as evidenced by the DSC values reported in Table 4.3. In addition, Figure 4.3 shows some representative samples of the region of the median nerve from predictions of the proposed model and of the U-Net based models. The U-net models often confounded the median nerve section with other rounded structures regardless of their shape or characteristic pattern. The Lightweight U-Net models, in particular, obtained the worst performances generating a lot of FP predictions, thus resulting not being very effective in median nerve localization. The proposed model, instead, incorrectly identifies only the infrequent morphologies, thus all images belonging to the same patients which present rare anatomical variants at the carpal tunnel level.

In a few cases, though, the developed Mask R-CNN didn't lead to a perfect segmentation, but even in such cases, it achieved better performances than the other models. As displayed in Fig. 4.3, the model struggles to interpret US images with relatively infrequent anatomical variants, like in contiguity with vessels as in Fig. 4.3c, and in the presence of bifid median nerve as in Fig. 4.3d.

In addition, poor definition of nerve borders, presence of multiple rounded hypoechoic areas, complex fascicular pattern typical of peripheral nerves, and inhomogeneities of the nerve section could contribute to making the detection harder. Results of the ablation studies reported in Table 4.2 highlighted how a deeper backbone granted good outcomes, and it could be appreciated that concatenation of several augmentations provides better results and more generalization than considering no augmentations on training data. In the future, it can be considered to introduce color augmentations, like brightness variation. As in Table 4.2, the increase in the output mask resolution from the segmentation head provided generally more accurate results. In fact, there are considerable improvements passing from 28x28 to 112x112 pixels output mask resolution, and lower performances are also visible in Mask56 compared to the proposed model. In addition, in Table 4.2 it is possible to appreciate that the concatenation of several augmentation generalized results better than considering single operations, like only rotation and only translation, on training data.

To increase the algorithm generalization, indeed, it is fundamental to expand the dataset with US images encompassing a wider spectrum of normal anatomy at the carpal tunnel level.

In future work, it could be interesting to consider pretraining on larger US existing datasets to improve model accuracy and reliability. The dataset should also be enlarged considering different US image acquisition equipment, lower-frequency probes and maybe involving more research centers in the study to strengthen generalizability further. It could be interesting even to approach the problem including different diagnostic tests and imaging the median nerve at the carpal tunnel from a different perspective and considering different wrist motion.

4.5 Conclusion

In this work, a DL approach was developed to provide a reliable tool for the automatic segmentation of the median nerve in US images, and from which directly measure the CSA of the median nerve. Even though improvements are needed to deploy the model in the clinical practice, the promising results obtained have shown the potentiality of such DL approach, which could allow to support beginner sonographers, to introduce standardized protocols, and thus to possibly support CTS diagnosis through US inspection.

In future, spatio-temporal information [252] could be included: other than improving median nerve segmentation, US videos also allow to evaluate an additional relevant parameter for CTS, the median nerve mobility. Distance-field regression for accurate nerve delineation could be investigated too, considering the promising results achieved in close fields [253]; alternatively, improving the detector of a Cascade Mask R-CNN as in [254] could be explored to minimize inaccurate localization and low recognition accuracy.

Chapter 5

Conclusion

Throughout these three years of doctoral research, a driving force and an enduring curiosity have steered every stage of my work: delving deeply into the benefits of employing AI within the realm of healthcare data archiving, with a particular focus on medical image analysis.

The main motivation for this thesis stems from the belief that integrating AI with digital medical archive systems can pave new pathways for early diagnosis and personalized care, in addition to improving the efficiency and accuracy of daily operations and resource management.

Digital medical archives, a vast repository of patient data, offer an unparalleled opportunity for DL-enhanced analysis of medical data, including unstructured data like medical images. However, the transition from theoretical or experimental DL models to practical, clinically impactful applications is still an open challenge. The main critical aspects of this transition are: i) the access to structured and secure medical archives; and ii) the robust validation of DL algorithms in diverse, real-world clinical settings. Having accessibility to medical digital archives and leveraging of the vast repository of diverse data and historical clinical scenarios, would be pivotal to produce accurate and generalizable DL algorithms. However, even considering a limited amount of medical images (either gathered for the research community or from a specific hospital), the potential of DL can still be recognized as a valuable and effective tool to support clinicians' decision-making process. Therefore, this thesis focused on the development of innovative DL methodologies specific for different tasks and different diagnostic purposes, based on different images acquired from various sources in the clinical practice and stored in dedicated digital archives.

With the goal of advancing the field, Chapter 2 presents an in-depth exploration of AI applications within laryngoscopy. The chapter aims to provide a dual-faceted contribution: firstly, by offering an automated, reliable, and objective assessment of VF motility, and secondly, by introducing an automated technique for segmenting tumors in the UADT. This dual approach underlines the utility of AI in minimizing diagnostic subjectivity and enhancing diagnostic accuracy.

Chapter 3 addresses the well-known problem of data scarcity and lack of annotated data by proposing a novel self-supervised DA approach for IVD segmentation in MRI images. The efficacy of this method is validated through extensive testing on three diverse and publicly accessible datasets within the field, showcasing the capability of the proposed technique to improve performance across varying image domains and acquisition protocols. This advancement permits the utilization of unlabeled data, potentially revolutionizing the approach to medical image segmentation.

In Chapter 4, the focus shifts to the estimation of the median nerve's CSA as a means to assess the progression of CTS in US images. The findings affirm the potential of DL algorithms to not only enhance the speed of diagnoses and alleviate the clinicians' workload, but also to facilitate the standardization of diagnostic practice, thereby enhancing the overall quality of patient care.

Overall, the described methodologies are designed to address the challenges associated with qualitative evaluations, high intra- and inter-observer variability, and the scarcity of extensive datasets. The ultimate goal is to aid medical professionals in clinical practice by leveraging the wealth of information contained in digital medical archives.

5.1 Thesis Contributions

The contribution of this PhD research can be summarized as follows:

- *Developing innovative and effective AI methods for analyzing endoscopic images of the larynx, with a particular focus on vocal cords.*

  Endoscopy is widely used as a diagnostic tool in a variety of medical contexts, providing real-time, high-resolution imaging. In clinical practice, endoscopic images are visually examined by medical personnel, which is time-consuming and characterized by high inter- and intra-rater variability [136]. Furthermore, endoscopic images show peculiar challenges including poor contrast, low signal-to-noise ratio, presence of motion blurring, and tissue motion. Thus, automatic and reliable endoscopic frame examination would improve diagnosis precision and reliability. In this context, DL has the potential to tackle the variability of endoscopic frames providing a more rigorous, objective, reliable, and repeatable perspective to the analysis. Chapter 2 introduces and discusses the most innovative and effective DL methods for the analysis of endoscopic images of the UADT, providing a concise yet comprehensive resource for emerging researchers, as well as a reference material for experienced professionals in the field. Delving deeper into the field of laryngoscopy, the Chapter provides two specific DL applications for the analysis of endoscopic images of the UADT. In Section 2.2, a novel heatmaps regression network is proposed to estimate VF motility, instead of using a direct regression approach. The method proposed is trained and evaluated using endoscopic frames from a dedicated archive. Experimental results showcase that heatmap regression yields highly accurate estimations of keypoint locations. The proposed end-to-end approach has moreover the potential for seamless integration into the device and practical adoption in the actual clinical practice. VF motility estimation is further investigated in Section 2.3, where starting from the coordinates of the five keypoints, clinically relevant features are handcrafted to train classification models and discriminate between VF preserved motility and fixation. Finally, being laryngoscopy the gold standard screening diagnostic tool for the diagnosis of precancerous lesions and early cancer of the larynx, in Sec. 2.4 a method for the automatic segmentation of UADT cancer is outlined. Utilizing a Mask R-CNN trained and evaluated on a dataset of endoscopic frames from a dedicated medical archive, the proposed method demonstrated effectiveness in 76.5% of images across three different anatomical sites: larynx/hypopharynx, oral cavity, and oropharynx. The promising results underscore the potential of DL algorithms for accurate instance segmentation in UADT diagnostics.

- *Evaluating the feasibility of contrastive learning to train a self-supervised DA model for IVD* *segmentation.*

For an extended period, the effectiveness of DL algorithms has heavily relied on the availability of high-quality, labeled data. This requirement poses a significant challenge in the training process, particularly in the field of medical imaging, where data labeling requires the expertise of trained professionals. Thus, a primary focus within the research community has been the development of self-learning mechanisms that may use unstructured data effectively. In addressing the aforementioned issue of labeling data, extensive research has been directed toward self-supervised learning techniques able to capture subtle nuances in data. Nonetheless, the direct application of these methods in the medical image domain, particularly for DA, represents a non-trivial challenge. Firstly, when considering data from different sources DL models exhibit performance limitations. This aspect, known as domain shift, becomes especially evident in the case of MRI, where different scanners and acquisition modalities produce very different images in terms of intensity distribution. In different fields of medical image analysis, various DA techniques have been used to address the challenges associated with transferring knowledge from a source domain to unlabelled target domains. Guided by these research needs, Chapter 3 investigates the use of contrastive learning, specifically performing pretext tasks on intensity features for self-supervised DA applied to a well-known problem like IVD segmentation in MRI, taking advantage of publicly available datasets of the field. The proposed approach involves the application of intensity pretext tasks to extract discriminative features. This aspect is of particular interest for MR images, for which variations in hardware and software create non-standard tissue intensities. This dual-task segmentation model based on intensity distortion pretext tasks, demonstrated robust performance across diverse datasets, thus overcoming challenges associated with intensity variations arising from different acquisition devices and medical conditions. By leveraging unlabeled data, the model achieves superior segmentation performance compared to a supervised segmentation model. The successful application of unsupervised DA in the context of medical image analysis is a significant contribution of this study, and the proposed approach holds the potential to be adapted to address other segmentation challenges.

- *Developing a new framework for median nerve segmentation for CTS estimation in US rheumatological images.*

  CTS is one of the most common rheumatic diseases, accounting for 90% of peripheral entrapment neuropathies, and affecting up to 5% of the general population [233]. In clinical practice, its diagnosis mainly relies on the clinical patient history and physical examination [235]. To introduce reliable measurements and make CTA assessment more standardized, US imaging can be used as it allows accurate identification of the median nerve structural morphology and anomalies [233]. Despite the growing interest in the potential of US for CTS assessment, its routine utilization faces several notable challenges, primarily related to the reliance on sonographer evaluation and the need for specialized expertise. Therefore, in Chapter 4 to reduce intra- and inter-clinician variability and to provide objective support for CTS evaluation, a DL framework is developed that demonstrates promise in its ability to localize and segment the median nerve effectively and facilitate automatic measurements of CSA, which can be invaluable in assessing CTS progression. By automating these tasks, the framework reduces the dependence on specialized sonographer skills, enhancing the consistency of carpal tunnel US evaluations across different practitioners. The significance of this work lies in its contribution to addressing the challenges associated with CTS assessment using US images. It presents a potential solution for the automatic localization and segmentation of the median nerve, ultimately improving the quality and standardization of diagnostic practices in CTS evaluation. This advancement opens up opportunities for further research in the field to streamline diagnostic processes.

5.2 Impact

DL technologies have the potential to outperform humans in some visual and auditory perception tasks. This superior performance indicates their significant potential for application in the field of medicine and healthcare. Particularly noteworthy is their use in the analysis of medical imaging, where DL can offer enhanced accuracy and efficiency. Each medical image domain is characterized by its unique challenges, including operator dependence, noise, artifacts, limited field of view, and variability across different acquisition devices. The outcomes of this PhD research have the potential to positively impact the way medical images stored in digital health archives are used to support the diagnostic process. The research conducted in these three years centers on the development of DL-based algorithms tailored to enhance diagnostic processes in specific medical contexts where clinical routines often lack standardized and uniform measurements. These innovative algorithms not only strengthen the reliability of diagnostic evaluations, effectively reducing variability and uncertainty but also facilitate the standardization of procedures and measurements within the medical field. By introducing these AI-driven tools into the diagnostic landscape, my research paves the way for more consistent, efficient, and accurate medical assessments, transforming clinical practices and elevating patient care standards. The implementation of these DL algorithms represents a major advancement in medical diagnostics, offering a novel paradigm where precision, reliability, and standardization constitute the base of patient evaluation and treatment planning.

5.3 Future perspectives

From the dissertation of this thesis, it is evident that DL has entered various application areas of medical image analysis. However, although DL techniques continuously update state-of-the-art performance results for various application domains, much remains to be done. It is known that the higher the data variability of the acquired dataset, the higher the performance improvement that can be achieved with DL. However, differently from other research areas in which datasets are large and publicly available (e.g., more than 1 million annotated multi-label natural images in ImageNet [63]), in the field of medical imaging there is an objective difficulty in retrieving structured data from medical archives along with clinicians annotations. Limited training data is a bottleneck for i) developing further clinical DL applications ii) integrating DL algorithms in clinical practice. Regardless, it is hard to expect that medical imaging datasets will reach the appropriate size in the short-term future, pointing out the need to find different DL training strategies, that can be suited for smaller databases [255]. Transfer learning has been widely exploited and its effectiveness is also clearly shown in Chapter 2 and Chapter 4 and several other works [256]. Semi- and weakly supervised learning have been successfully used [257], but their application is more suitable when larger sets of unlabeled data are available. In addition to the aforementioned techniques, other approaches that address small annotated databases are the unsupervised learning ones, which exploit data reconstruction while learning a good representation, or learning similarities among data employing contrastive learning loss functions. Recent progress on self-supervised visual representations has paved the way for applying these methodologies in DA applications [208, 258, 259], as also demonstrated in Chapter 3. These techniques may be exploited to integrate different datasets of the field, increasing the patient population representation, as well as acquisition modalities, thus reducing the possibility of generating biased results. Federated learning is a novel paradigm for data-private multi-institutional collaborations that could also be exploited to incentivize collaboration among clinical centers. It consists of the training of algorithms across multiple decentralized edge devices without exchanging private data.

When considering the prospects of AI in healthcare, the development of more accurate and generalizable algorithms that can truly support diagnosis hinges upon a critical factor: accessibility to the vast reservoir of historical data stored within medical digital archives. However, this aspiration comes bundled with a complex web of privacy and ethical concerns that require meticulous consideration and proactive solutions. The sensitive nature of patient information requires stringent safeguards against unauthorized access and thus, data accessibility strictly requires privacy protection strategies, such as robust data de-identification and anonymization techniques. Moreover, AI systems must face ethical dilemmas that may arise and grapple with the challenge of mitigating biases: for instance, ensuring equitable representation of all demographic groups to prevent disparities in diagnosis; or recognizing and accounting for prevalent medical conditions to address the possibility of inadvertent misrepresentation of a minority group, resulting in under- or misdiagnosis of certain medical conditions, potentially perpetuating health disparities.

To navigate this intricate landscape, it is essential to establish a comprehensive framework for AI governance within medical digital archives. The core principles of this AI governance should be inclusivity, transparency, ethics, and privacy, inspired by consolidated practices among archivists as well as in social scientists, historians, and anthropologists. Thus, unlocking the potential of digital medical image archives requires not only interdisciplinary collaboration between clinicians, technologists, and IT professionals but also rigorous adherence to ethical principles.

Bibliography

[1] Michael Moss, David Thomas, and Tim Gollins. The reconfiguration of the archive as data to be mined. *Archivaria*, 86:118–151, Nov. 2018.

[2] Daniel Burda and Teuteberg Frank. Sustaining accessibility of information through digital preservation: A literature review. *Journal of Information Science*, 39:442 – 458, 2013.

[3] Wei Guo, Yunyi Fang, Weimei Pan, and Dekun Li. Archives as a trusted third party in maintaining and preserving digital records in the cloud environment. *Records Management Journal*, 26:170–184, 2016.

[4] Gi Oliver, Bi Chawner, and Hai Ping Liu. Implementing digital archives: issues of trust. *Archival Science*, 11:311–327, 2011.

[5] Miguel Costa, Daniel Gomes, and Mário J. Silva. The evolution of web archiving. *International Journal on Digital Libraries*, 18:191–205, 2017.

[6] Jonathan Stray. Making artificial intelligence work for investigative journalism. *Digital Journalism*, 7:1076–1097, 2019.

[7] Amber L. Cushing and Giulia Osti. "So how do we balance all of these needs?": How the concept of AI technology impacts digital archival expertise. *Journal of Documentation*, pages 12–29, 2022.

[8] Lise Jaillant and Annalina Caputo. Unlocking digital archives: Cross-disciplinary perspectives on ai and born-digital data. *AI & SOCIETY*, 37(3):823–835, 2022.

[9] Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. Archival records and training in the age of big data. *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, 44B:179–199, 2018.

[10] Rolan Gregory, Humphries Glen, Jeffrey Lisa, Samaras Evanthia, Antsoupova Tatiana, and Stuart Katharine. More human than human? Artificial intelligence in the archive. *Archives & Manuscripts*, 47(2):179–203, Nov. 2018.

[11] Kenneth Thibodeau. Breaking down the invisible wall to enrich archival science and practice. *2016 IEEE International Conference on Big Data (Big Data)*, pages 3277–3282, 2016.

[12] Chenhao Su, Sheng Gao, and Si Li. Gate: Graph-attention augmented temporal neural network for medication recommendation. *IEEE Access*, 8:125447–125458, 2020.

[13] Christopher A. Lee. Computer-assisted appraisal and selection of archival materials. *2018 IEEE International Conference on Big Data (Big Data)*, pages 2721–2724, 2018.

[14] Adrian Cosma, Mihai Ghidoveanu, Michael Panaitescu-Liess, and Marius Claudiu Popescu. Self-supervised representation learning on document images. *ArXiv*, abs/2004.10605, 2020.

[15] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, R. Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2021.

[16] Jason R. Baron and Nathaniel Payne. Dark archives and edemocracy: Strategies for overcoming access barriers to the public record archives of the future. In *2017 Conference for E-Democracy and Open Government (CeDEM)*, pages 3–11, 2017.

[17] Tim Hutchinson. Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections. *2017 IEEE International Conference on Big Data (Big Data)*, 2017.

[18] Tim Hutchinson. Protecting privacy in the archives: Supervised machine learning and born-digital records. *2018 IEEE International Conference on Big Data (Big Data)*, 2018.

[19] Ross Spencer. Binary trees? automatically identifying the links between born-digital records. *Archives and Manuscripts*, 45(2):77–99, 2017.

[20] Woojin Paik, Sibel Yilmazel, Eric Brown, Maryjane Poulin, Stephane Dubon, and Christophe Amice. Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In *International Conference on Knowledge Capture*, pages 116–122, 2001.

[21] PierLuigi Spinosa, Gerardo Giardiello, Manola Cherubini, Simone Marchi, Giulia Venturi, and Simonetta Montemagni. Nlp-based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, page 40–49. Association for Computing Machinery, 2009.

[22] Rob Knapen, Thomas Hüsing, Klaus Jacob, Yke van Randen, Stefan Reis, Onno Roosenschoon, and Sander Janssen. Metadata extraction using semantic and natural language processing techniques. *Proceedings of the International Environmental Modelling and Software Society (iEMSs) 7th Intl. Congress on Env. Modelling and Software*, pages 385–391, 2014.

[23] Matthew J. Connelly, Raymond Hicks, Robert Jervis, Arthur Spirling, and Clara H. Suong. Diplomatic documents data for international relations: the freedom of information archive database. *Conflict Management and Peace Science*, 38(6):762–781, 2021.

[24] B. Lee. Machine learning, template matching, and the international tracing service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans. *Digit. Scholarsh. Humanit.*, 34:513–535, 2018.

[25] Mark Bell. From tree to network: reordering an archival catalogue. *Records Management Journal*, 30:379–394, 2020.

[26] Thomas Sødring, Petter Reinholdtsen, and David Massey. A record-keeping approach to managing iot-data for government agencies. *Records Management Journal*, 30(2):221–239, 2020.

[27] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 306–316. Association for Computing Machinery, 2020.

[28] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684, 2020.

[29] Lorraine Dong, Polina Ilieva, and Aimee Medeiros. Data dreams: Planning for the future of historical medical documents. *Journal of the Medical Library Association*, 106(4), 2018.

[30] Guthrie S. Birkhead, Michael Klompas, and Nirav R. Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual Review of Public Health*, 36(1):345–359, 2015.

[31] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: Data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1 – 5, 2010.

[32] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[33] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018.

[34] Suraj Tekchandani, Jigar Shah, and Archana Singh. Archive system using big data for health care: Analysis, architecture, and implementation. In *Data Science and Intelligent Applications*, pages 1–11, Singapore, 2021. Springer Singapore.

[35] Nicholas Ayache. Medical imaging in the age of artificial intelligence. *Healthcare and Artificial Intelligence*, pages 89–91, 2020.

[36] Rogier Van de Wetering and Ronald Batenburg. A PACS maturity model: A systematic meta-analytic review on maturation and evolvability of pacs in the hospital enterprise. *International journal of medical informatics*, 78 2:127–40, 2009.

[37] Seong K Mun, Kenneth H Wong, Shih-Chung B Lo, Yanni Li, and Shijir Bayarsaikhan. Artificial intelligence for the future radiology diagnostic service. *Frontiers in molecular biosciences*, 7:614258, 2021.

[38] Jorge Miguel Silva, Eduardo Pinho, E Monteiro, J. F. Silva, and C Costa. Controlled searching in reversibly de-identified medical imaging archives. *Journal of biomedical informatics*, 77:81–90, 2018.

[39] Aren Shah, Pooja Sai Muddana, and Safwan Halabi. A review of core concepts of imaging informatics. *Cureus*, 2022.

[40] Michele Larobina and Loredana Murino. Medical image file formats. *Journal of digital imaging*, 27:200–206, 2014.

[41] Andreas S Panayides, Amir Amini, Nenad D Filipovic, Ashish Sharma, Sotirios A Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina Nikita, Ben Veasey, Michalis Zervakis, Joel Saltz, and Constantinos Pattichis. AI in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and health informatics*, 24:1837–1857, 2020.

[42] Hisatomi Arima. Utilizing big data for public health. *Journal of Epidemiology*, 26:105–105, 2016.

[43] Elmar Kotter. Basic workflow of medical imaging. In *Basic Knowledge of Medical Imaging Informatics: Undergraduate Level and Level I*, pages 41–53. Springer, 2021.

[44] Florian Schwind, H. Münch, A. Schröter, R. Brandner, U. Kutscha, A. Brandner, O. Heinze, B. Bergh, and U. Engelmann. Long-term experience with setup and implementation of an ihe-based image management and distribution system in intersectoral clinical routine. *International Journal of Computer Assisted Radiology and Surgery*, 13:1727–1739, 2018.

[45] Ken Chang, N. Balachandar, Carson K. Lam, Darvin Yi, James M. Brown, Andrew Beers, B. Rosen, D. Rubin, and Jayashree Kalpathy-Cramer. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association : JAMIA*, 25:945 – 954, 2018.

[46] European Society of Radiology (ESR). The new EU general data protection regulation: what the radiologist should know. *Insights into Imaging*, 8(3):295–299, 2017.

[47] Eliot L. Siegel and David S. Channin. Integrating the healthcare enterprise: a primer. part 1. introduction. *Radiographics: a review publication of the Radiological Society of North America, Inc*, 21 5:1339–1341, 2001.

[48] David S. Channin. Integrating the healthcare enterprise: a primer. part 2. seven brides for seven brothers: the ihe integration profiles. *Radiographics: a review publication of the Radiological Society of North America, Inc*, 21 5:1343–1350, 2001.

[49] David S. Channin, C Parisot, V Wanchoo, A Leontiev, and EL. Siegel. Integrating the healthcare enterprise: a primer: Part 3. what does ihe do for me? *Radiographics: a review publication of the Radiological Society of North America, Inc*, 21 5:1351–1358, 2001.

[50] Rita Noumeir, Alain Lemay, and Jean-Marc Lina. Pseudonymization of radiology data for research purposes. *Journal of digital imaging*, 20:284–295, 2007.

[51] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[52] Oliver Diaz, Kaisar Kushibar, Richard Osuala, Akis Linardos, Lidia Garrucho, Laura Igual, Petia Radeva, Fred Prior, Polyxeni Gkontra, and Karim Lekadir. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica Medica*, 83:25–37, 2021.

[53] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[54] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017.

[55] Puneet Sharma, Michael Suehling, Thomas Flohr, and Dorin Comaniciu. Artificial intelligence in diagnostic imaging: status quo, challenges, and future opportunities. *Journal of thoracic imaging*, 35:S11–S16, 2020.

[56] DM Anisuzzaman, Chuanbo Wang, Behrouz Rostami, Sandeep Gopalakrishnan, Jeffrey Niezgoda, and Zeyun Yu. Image-based artificial intelligence in wound assessment: A systematic review. *Advances in Wound Care*, 11(12):687–709, 2022.

[57] Ravi Manne and Sneha C Kantheti. Application of artificial intelligence in healthcare: chances and challenges. *Current Journal of Applied Science and Technology*, 40(6):78–89, 2021.

[58] Martina Gurgitano, Salvatore Alessio Angileri, Giovanni Maria Rodà, Alessandro Liguori, Marco Pandolfi, Anna Maria Ierardi, Bradford J Wood, and Gianpaolo Carrafiello. Interventional radiology ex-machina: Impact of artificial intelligence on practice. *La radiologia medica*, 126:998–1006, 2021.

[59] K He, X Zhang, S Ren, et al. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[60] Huanhuan Zhang and Yufei Qie. Applying deep learning to medical imaging: A review. *Applied Sciences*, 13(18), 2023.

[61] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[62] Rehan Ashraf, Muhammad Asif Habib, M. Akram, M. Latif, M. S. A. Malik, M. Awais, Saadat Hanif Dar, Toqeer Mahmood, Muhammad Yasir, and Zahoor Abbas. Deep convolution neural network for big data medical image classification. *IEEE Access*, 8:105659–105670, 2020.

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[64] 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, 2016.

[65] Tom Brosch and Roger Tam. Manifold learning of brain mris by deep learning. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 633–640. Springer Berlin Heidelberg, 2013.

[66] Sergey M. Plis, Devon R. Hjelm, Ruslan Salakhutdinov, Elena A. Allen, Henry J. Bockholt, Jeffrey D. Long, Hans J. Johnson, Jane S. Paulsen, Jessica A. Turner, and Vince D. Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience*, 8, 2014.

[67] Maria Chiara Fiorentino, Francesca Pia Villani, Mariachiara Di Cosmo, Emanuele Frontoni, and Sara Moccia. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, 2022.

[68] Haimiao Zhang and Bin Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*, 8:311–340, 2020.

[69] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *J Big Data*, 60(6), 2019.

[70] Hu Chen, Yi Zhang, Weihua Zhang, Peixi Liao, Ke Li, Jiliu Zhou, and Ge Wang. Low-dose ct via convolutional neural network. *Biomedical Optics Express*, 8:679–694, 01 2017.

[71] Y Han, J Yoo, HH Kim, HJ Shin, K Sung, and JC. Ye. Deep learning with domain adaptation for accelerated projection-reconstruction mr. *Magn Reson Med.*, 80(3):1189–1205, 2018.

[72] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI.*, 2015.

[73] Fausto Milletarì, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[74] Yuanpu Xie, Zizhao Zhang, Manish Sapkota, and Lin Yang. Spatial clockwork recurrent neural network for muscle perimysium segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 185–193. Springer International Publishing, 2016.

[75] Pim Moeskops, Jelmer M. Wolterink, Bas H. M. van der Velden, Kenneth G. A. Gilhuijs, Tim Leiner, Max A. Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 478–486. Springer International Publishing, 2016.

[76] Mahsa Shakeri, Stavros Tsogkas, Enzo Ferrante, Sarah Lippe, Samuel Kadoury, Nikos Paragios, and Iasonas Kokkinos. Sub-cortical brain structure segmentation using f-cnn's. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 269–272, 2016.

[77] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nat Rev Cancer*, 18:500–510, 2018.

[78] Yunliang Cai, Mark Landis, David T. Laidley, Anat Kornecki, Andrea Lum, and Shuo Li. Multi-modal vertebrae recognition using transformed deep convolution network. *Computerized Medical Imaging and Graphics*, 51:11–19, 2016.

[79] Ashnil Kumar, Pradeeba Sridar, Ann Quinton, R. Krishna Kumar, Dagan Feng, Ralph Nanan, and Jinman Kim. Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794, 2016.

[80] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using cnns. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 230–238. Springer International Publishing, 2016.

[81] Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Sandra Smith, Bernhard Kainz, and Daniel Rueckert. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 203–211. Springer International Publishing, 2016.

[82] Bin Kong, Yiqiang Zhan, Min Shin, Thomas Denny, and Shaoting Zhang. Recognizing end-diastole and end-systole frames via deep temporal regression network. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, pages 264–272. Springer International Publishing, 2016.

[83] S.-C.B. Lo, S.-L.A. Lou, Jyh-Shyan Lin, M.T. Freedman, M.V. Chien, and S.K. Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[84] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 239–246. Springer International Publishing, 2016.

[85] Qi Dou, Hao Chen, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567, 2017.

[86] Samaneh Abbasi, Meysam Tavakoli, Hamid Reza Boveiri, Mohammad Amin Mosleh Shirazi, Raouf Khayami, Hedieh Khorasani, Reza Javidan, and Alireza Mehdizadeh. Medical image registration using unsupervised deep neural network: A scoping literature review. *Biomedical Signal Processing and Control*, 73:103444, 2022.

[87] Axel Boese, Cora Wex, Roland Croner, Uwe Bernd Liehr, Johann Jakob Wendler, Jochen Weigt, Thorsten Walles, Ulrich Vorwerk, Christoph Hubertus Lohmann, Michael Friebe, et al. Endoscopic imaging technology today. *Diagnostics*, 12(5):1262, 2022.

[88] Alberto Paderno, F. Christopher Holsinger, and Cesare Piazza. Videomics: bringing deep learning to diagnostic endoscopy. *Curr Opin Otolaryngol Head Neck Surg*, 29:143–148, 2021.

[89] Sharib Ali, Barbara Braden, Dominique Lamarque, Stefano Realdon, Adam Bailey, Renato Cannizzaro, Noha Ghatwary, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection and segmentation (edd2020). *IEEE DataPort*, 2020.

[90] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical image analysis*, 71:102058, 2021.

[91] Yutaka Saito, Shinya Kodashima, Takahisa Matsuda, Koji Matsuda, Mitsuhiro Fujishiro, Kiyohito Tanaka, Kiyonori Kobayashi, Chikatoshi Katada, Takahiro Horimatsu, Manabu Muto, et al. Current status of diagnostic and therapeutic colonoscopy in japan: The japan endoscopic database project. *Digestive Endoscopy*, 34(1):144–152, 2022.

[92] Yutaka Okagawa, Seiichiro Abe, Masayoshi Yamada, Ichiro Oda, and Yutaka Saito. Artificial intelligence in endoscopy. *Digestive Diseases and Sciences*, 67(5):1553–1572, 2022.

[93] Jamie Catlow, Benjamin Bray, Eva Morris, and Matt Rutter. Power of big data to improve patient care in gastroenterology. *Frontline Gastroenterology*, 13(3):237–244, 2022.

[94] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical Image Analysis*, 76:102306, 2022.

[95] Cesare Piazza, Francesca D. Bon, Giorgio Peretti, and Piero Nicolai. 'biologic endoscopy': Optimization of upper aerodigestive tract cancer evaluation. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 19(2):67–76, 2011.

[96] Pablo Gómez, Andreas M. Kist, Patrick Schlegel, David A. Berry, Dinesh K Chhetri, Stephan Dürr, Matthias Echternach, Aaron M. Johnson, Stefan Kniesburges, Melda Kunduk, Youri Maryn, Anne Schützenberger, Monique Verguts, and Michael Döllinger. BAGLS, a multihospital benchmark for automatic glottis segmentation. science data, 7, article id 186, 2020.

[97] Khalid Raza and Nripendra K Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9):1059–1077, 2021.

[98] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.

[99] Ilaria Patrini, Michela Ruperti, Sara Moccia, Leonardo S Mattos, Emanuele Frontoni, and Elena De Momi. Transfer learning for informative-frame selection in laryngoscopic videos through learned features. *Medical & Biological Engineering & Computing*, 58(6):1225–1238, 2020.

[100] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

[101] Yurong He, Yingduan Cheng, Zhigang Huang, Wen Xu, Rong Hu, Liyu Cheng, Shizhi He, Changli Yue, Gang Qin, Yan Wang, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. *Annals of Translational Medicine*, 9(24), 2021.

[102] Nazila Esmaeili, Esam Sharaf, Elmer Jeto Gomes Ataide, Alfredo Illanes, Axel Boese, Nikolaos Davaris, Christoph Arens, Nassir Navab, and Michael Friebe. Deep convolution neural network for laryngeal cancer classification on contact endoscopy-narrow band imaging. *Sensors*, 21:8157, 2021.

[103] Qian Zhao, Yuqing He, Yanda Wu, Dongyan Huang, Yang Wang, Cai Sun, Jun Ju, Jiasen Wang, and Jeremy Jianshuo-li Mahr. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Medical Physics*, 49(1):432–442, 2022.

[104] Chang-Chiun Huang, Yi-Shing Leu, Chung-Feng Jeffrey Kuo, Wen-Lin Chu, Yueng-Hsiang Chu, and Han-Cheng Wu. Automatic recognizing of vocal fold disorders from glottis images. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 228(9):952–961, 2014.

[105] Michael E Dunham, Keonho A Kong, Andrew J McWhorter, and Lacey K Adkins. Optical biopsy: automated classification of airway endoscopic findings using a convolutional neural network. *The Laryngoscope*, 132:S1–S8, 2022.

[106] Cesare Piazza, Francesca Del Bon, Alberto Paderno, Paola Grazioli, Pietro Perotti, Diego Barbieri, Alessandra Majorana, Elena Bardellini, Giorgio Peretti, and Piero Nicolai. The diagnostic value of narrow band imaging in different oral and oropharyngeal subsites. *European Archives of Oto-Rhino-Laryngology*, 273:3347–3353, 2016.

[107] Shamik Mascharak, Brandon J Baird, and F Christopher Holsinger. Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning. *The Laryngoscope*, 128(11):2514–2520, 2018.

[108] Bofan Song, Sumsum Sunny, Ross D Uthoff, Sanjana Patrick, Amritha Suresh, Trupti Kolur, G Keerthi, Afarin Anbarani, Petra Wilder-Smith, Moni Abraham Kuriakose, et al. Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. *Biomedical optics express*, 9(11):5318–5329, 2018.

[109] Atsushi Inaba, Keisuke Hori, Yusuke Yoda, Hiroaki Ikematsu, Hiroaki Takano, Hiroki Matsuzaki, Yoshiki Watanabe, Nobuyoshi Takeshita, Toshifumi Tomioka, Genichiro Ishii, et al. Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning. *Head & Neck*, 42(9):2581–2592, 2020.

[110] Hao Xiong, Peiliang Lin, Jin-Gang Yu, Jin Ye, Lichao Xiao, Yuan Tao, Zebin Jiang, Wei Lin, Mingyue Liu, Jingjing Xu, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*, 48:92–99, 2019.

[111] Clyde Matava, Evelina Pankiv, Sam Raisbeck, Monica Caldeira, and Fahad Alam. A convolutional neural network for real time classification, identification, and labelling of vocal cord and tracheal using laryngoscopy and bronchoscopy video. *Journal of medical systems*, 44:1–10, 2020.

[112] Muhammad Adeel Azam, Claudio Sampieri, Alessandro Ioppi, Stefano Africano, Alberto Vallin, Davide Mocellin, Marco Fragale, Luca Guastini, Sara Moccia, Cesare Piazza, et al. Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. *The Laryngoscope*, 132(9):1798–1806, 2022.

[113] Max-Heinrich Laves, Jens Bicker, Lüder A Kahrs, and Tobias Ortmaier. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International journal of computer assisted radiology and surgery*, 14:483–492, 2019.

[114] Alberto Paderno, Cesare Piazza, Francesca Del Bon, Davide Lancini, Stefano Tanagli, Alberto Deganello, Giorgio Peretti, Elena De Momi, Ilaria Patrini, Michela Ruperti, et al. Deep learning for automatic segmentation of oral and oropharyngeal cancer using narrow band imaging: preliminary experience in a clinical perspective. *Frontiers in Oncology*, 11:626602, 2021.

[115] Mona Kirstin Fehling, Fabian Grosch, Maria Elke Schuster, Bernhard Schick, and Jörg Lohscheller. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network. *Plos one*, 15(2):e0227791, 2020.

[116] Chaofeng Li, Bingzhong Jing, Liangru Ke, Bin Li, Weixiong Xia, Caisheng He, Chaonan Qian, Chong Zhao, Haiqiang Mai, Mingyuan Chen, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. *Cancer Communications*, 38(1):1–11, 2018.

[117] Cesare Piazza, Stefano Mangili, Francesca Del Bon, Francesca Gritti, Claudia Manfredi, Piero Nicolai, and Giorgio Peretti. Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study. *European archives of oto-rhino-laryngology*, 269:207–212, 2012.

[118] Andreas M Kist, Julian Zilker, Pablo Gómez, Anne Schützenberger, and Michael Döllinger. Rethinking glottal midline detection. *Scientific reports*, 10(1):20723, 2020.

[119] Patrick Schlegel, Stefan Kniesburges, Stephan Dürr, Anne Schützenberger, and Michael Döllinger. Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings. *scientific reports*, 10(1):10517, 2020.

[120] Andreas M Kist, Pablo Gómez, Denis Dubrovskiy, Patrick Schlegel, Melda Kunduk, Matthias Echternach, Rita Patel, Marion Semmler, Christopher Bohr, Stephan Dürr, et al. A deep learning enhanced novel software tool for laryngeal dynamics analysis. *Journal of Speech, Language, and Hearing Research*, 64(6):1889–1903, 2021.

[121] Cesare Piazza, Nausica Montalto, Alberto Paderno, Valentina Taglietti, and Piero Nicolai. Is it time to incorporate 'depth of infiltration'in the t staging of oral tongue and floor of mouth cancer? *Current opinion in otolaryngology & head and neck surgery*, 22(2):81–89, 2014.

[122] Hong Jin Yoon, Seunghyup Kim, Jie-Hyun Kim, Ji-Soo Keum, Sang-Il Oh, Junik Jo, Jaeyoung Chun, Young Hoon Youn, Hyojin Park, In Gyu Kwon, et al. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. *Journal of clinical medicine*, 8(9):1310, 2019.

[123] Hiroko Nakahira, Ryu Ishihara, Kazuharu Aoyama, Mitsuhiro Kono, Hiromu Fukuda, Yusaku Shimamoto, Kentaro Nakagawa, Masayasu Ohmori, Taro Iwatsubo, Hiroyoshi Iwagami, et al. Stratification of gastric cancer risk using a deep neural network. *JGH Open*, 4(3):466–471, 2020.

[124] Mahul B Amin, Stephen B Edge, Frederick L Greene, David R Byrd, Robert K Brookland, Mary Kay Washington, Jeffrey E Gershenwald, Carolyn C Compton, Kenneth R Hess, Daniel C Sullivan, et al. *AJCC cancer staging manual*, volume 1024. Springer, 2017.

[125] Nat Adamian, Matthew R Naunheim, and Nate Jowett. An open-source computer vision tool for automated vocal fold tracking from videoendoscopy. *The Laryngoscope*, 131(1):E219–E225, 2021.

[126] Martin Curry, Anand Malpani, Ryan Li, Thomas Tantillo, Amod Jog, Ray Blanco, Patrick K Ha, Joseph Califano, Rajesh Kumar, and Jeremy Richmon. Objective assessment in residency-based training for transoral robotic surgery. *The Laryngoscope*, 122(10):2184–2192, 2012.

[127] Alberto Paderno, Alberto Deganello, Davide Lancini, and Cesare Piazza. Is the exoscope ready to replace the operative microscope in transoral surgery? *Current Opinion in Otolaryngology & Head and Neck Surgery*, 30(2):79–86, 2022.

[128] Inês Domingues, Gisèle Pereira, Pedro Martins, Hugo Duarte, João Santos, and Pedro Henriques Abreu. Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet. *Artificial Intelligence Review*, 53(6):4093–4160, 2020.

[129] Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.

[130] Alessandro Repici, Matteo Badalamenti, Roberta Maselli, Loredana Correale, Franco Radaelli, Emanuele Rondonotti, Elisa Ferrara, Marco Spadaccini, Asma Alkandari, Alessandro Fugazza, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology*, 159(2):512–520, 2020.

[131] Alberto Paderno, Francesca Gennaini, Alessandra Sordi, Claudia Montenegro, Davide Lancini, Francesca Pia Villani, Sara Moccia, and Cesare Piazza. Artificial intelligence in clinical endoscopy: insights in the field of videomics. *Frontiers in Surgery*, page 1361, 2022.

[132] Tiffany V Wang and Phillip C Song. Neurological voice disorders: a review. *International Journal of Head and Neck Surgery*, 13(1):32–40, 2022.

[133] J K R Menon, R M Nair, and S Priyanka. Unilateral vocal fold paralysis: can laryngoscopy predict recovery? a prospective study. *The Journal of Laryngology & Otology*, 128(12):1095–1104, 2014.

[134] Seth H. Dailey, James B. Kobler, Robert E. Hillman, Kittisard Tangrom, Ekawudh Thananart, Marcelo Mauri, and Steven M. Zeitels. Endoscopic measurement of vocal fold movement during adduction and abduction. *The Laryngoscope*, 115(1):178–183, 2005.

[135] Chloe Walton, Paul Carding, Erin Conway, Kieran Flanagan, and Helen Blackshaw. Voice outcome measures for adult patients with unilateral vocal fold paralysis: A systematic review. *The Laryngoscope*, 129(1):187–197, 2019.

[136] Daniel Voigt, Michael Döllinger, Anxiong Yang, Ulrich Eysholdt, and Jörg Lohscheller. Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. *Computer Methods and Programs in Biomedicine*, 99(3):275–288, 2010.

[137] Seth H. Dailey, James B. Kobler, Robert E. Hillman, Kittisard Tangrom, Ekawudh Thananart, Marcelo Mauri, and Steven M. Zeitels. Endoscopic measurement of vocal fold movement during adduction and abduction. *The Laryngoscope*, 115(1):178–183, 2005.

[138] Deanna Britton, Kathryn M. Yorkston, Tanya Eadie, Cara E. Stepp, Marcia A. Ciol, Carolyn Baylor, and Albert L. Merati. Endoscopic assessment of vocal fold movements during cough. *Annals of Otology, Rhinology & Laryngology*, 121(1):21–27, 2012.

[139] S T Kuna and C R Vanoye. Laryngeal response during forced vital capacity maneuvers in normal adult humans. *American Journal of Respiratory and Critical Care Medicine*, 150(3):729–734, 1994.

[140] J. Tod Olin, Matthew S. Clary, Dan Connors, Jordan Abbott, Susan Brugman, Yiming Deng, Xiaoye Chen, and Mark Courey. Glottic configuration in patients with exercise-induced stridor: A new paradigm. *The Laryngoscope*, 124(11):2568–2573, 2014.

[141] Manuel Diaz-Cadiz, Victoria S. McKenna, Jennifer M. Vojtech, and Cara E. Stepp. Adductory vocal fold kinematic trajectories during conventional versus high-speed videoendoscopy. *Journal of Speech, Language, and Hearing Research*, 62(6):1685–1706, 2019.

[142] Ali S. Hamad, Megan M Haney, Teresa E. Lever, and Filiz Bunyak. Automated segmentation of the vocal folds in laryngeal endoscopy videos using deep convolutional regression networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 140–148, 2019.

[143] Ahmed M. Yousef, Dimitar D. Deliyski, Stephanie R. Zacharias, Alessandro de Alarcon, Robert F. Orlikoff, and Maryam Naghibolhosseini. A deep learning approach for quantifying vocal fold dynamics during connected speech using laryngeal high-speed videoendoscopy. *Journal of Speech, Language, and Hearing Research*, 65(6):2098–2113, 2022.

[144] Gustavo Andrade-Miranda, Yannis Stylianou, Dimitar D. Deliyski, Juan Ignacio Godino-Llorente, and Nathalie Henrich Bernardoni. Laryngeal image processing of vocal folds motion. *Applied Sciences*, 10(5), 2020.

[145] Jianyu Lin, Emil S. Walsted, Vibeke Backer, James H. Hull, and Daniel S. Elson. Quantification and analysis of laryngeal closure from endoscopic videos. *IEEE Transactions on Biomedical Engineering*, 66(4):1127–1136, 2019.

[146] Nat Adamian, Matthew R. Naunheim, and Nate Jowett. An open-source computer vision tool for automated vocal fold tracking from videoendoscopy. *The Laryngoscope*, 131(1):E219–E225, 2021.

[147] Tiffany V. Wang, Nat Adamian, Phillip C. Song, Ramon A. Franco, Molly N. Huston, Nate Jowett, and Matthew R. Naunheim. Application of a computer vision tool for automated glottic tracking to vocal fold paralysis patients. *Otolaryngology–Head and Neck Surgery*, 165(4):556–562, 2021.

[148] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Computer Vision – ECCV 2016*, pages 717–732. Springer International Publishing, 2016.

[149] Sara Moccia, Lucia Migliorelli, Virgilio Carnielli, and Emanuele Frontoni. Preterm infants' pose estimation with spatio-temporal features. *IEEE Transactions on Biomedical Engineering*, 67(8):2370–2380, 2020.

[150] Francesca Pia Villani, Alberto Paderno, Maria Chiara Fiorentino, Alessandro Casella, Cesare Piazza, and Sara Moccia. Classifying vocal folds fixation from endoscopic videos with machine learning. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, 2023.

[151] Sara Moccia, Elena De Momi, Marco Guarnaschelli, Matteo Savazzi, Andrea Laborai, Luca Guastini, Giorgio Peretti, and Leonardo S Mattos. Confident texture-based laryngeal tissue classification for early stage diagnosis support. *Journal of Medical Imaging*, 4(3):034502, 2017.

[152] Ahmed M. Yousef, Dimitar D. Deliyski, Stephanie R.C. Zacharias, and Maryam Naghibolhosseini. Deep-learning-based representation of vocal fold dynamics in adductor spasmodic dysphonia during connected speech in high-speed videoendoscopy. *Journal of Voice*, 2022.

[153] Christian T. Herbst, Jakob Unger, Hanspeter Herzel, Jan G. Švec, and Jörg Lohscheller. Phasegram analysis of vocal fold vibration documented with laryngeal high-speed video endoscopy. *Journal of Voice*, 30(6), 2016.

[154] Jörg Lohscheller. Towards evidence based diagnosis of voice disorders using phonovibrograms. In *2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, pages 1–4, 2009.

[155] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[156] Vinayak Singh, Mahendra Kumar Gourisaria, and Himansu Das. Performance analysis of machine learning algorithms for prediction of liver disease. In *2021 IEEE 4th International*

Conference on Computing, Power and Communication Technologies (GUCON), pages 1–7, 2021.

[157] Md Abu Rumman Refat, Md. Al Amin, Chetna Kaushal, Mst Nilufa Yeasmin, and Md Khairul Islam. A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 654–659, 2021.

[158] Emanuele Colleoni, Sara Moccia, Xiaofei Du, Elena De Momi, and Danail Stoyanov. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robotics and Automation Letters*, 4(3):2714–2721, 2019.

[159] Sara Moccia, Lucia Migliorelli, Virgilio Carnielli, and Emanuele Frontoni. Preterm infants' pose estimation with spatio-temporal features. *IEEE Transactions on Biomedical Engineering*, 67(8):2370–2380, 2020.

[160] Alberto Paderno, Francesca Pia Pia Villani, Milena Fior, Giulia Berretti, Francesca Gennarini, Gabriele Zigliani, Emanuela Ulaj, Claudia Montenegro, Alessandra Sordi, Claudio Sampieri, Giorgio Peretti, Sara Moccia, and Cesare Piazza. Instance segmentation of upper aerodigestive tract cancer: site-specific outcomes. *Acta Otorhinolaryngologica Italica*, 43(4):283–290, 2023.

[161] Muhammad Adeel Azam, Claudio Sampieri, Alessandro Ioppi, Pietro Benzi, Giorgio Gregory Giordano, Marta De Vecchi, Valentina Campagnari, Shunlei Li, Luca Guastini, Alberto Paderno, Sara Moccia, Cesare Piazza, Leonardo Mattos, and Giorgio Peretti. Videomics of the upper aero-digestive tract cancer: Deep learning applied to white light and narrow band imaging for automatic segmentation of endoscopic images. front. *Frontiers in Oncology*, 12:900451, 2022.

[162] P Nogal, M Buchwald, M Staśkiewicz, et al. Endoluminal larynx anatomy model – towards facilitating deep learning and defining standards for medical images evaluation with artificial intelligence algorithms. *Otolaryngol Pol*, 76:1–9, 2022.

[163] K He, G Gkioxari, P Dollár, et al. Mask r-cnn. *IEEE Trans Pattern Anal Mach Intell*, 42:386–397, 2020.

[164] WK Cho, YJ Lee, HA Joo, et al. Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system. *Laryngoscope*, 131(11):2558–2566, 2021.

[165] BC Russell, A Torralba, KP Murphy, et al. Labelme: a database and web-based tool for image annotation. *Int J Comput Vis*, 77:57–173, 2008.

[166] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[167] T Lin, P Dollár, RB Girshick, et al. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[168] T Lin, M Maire, S Belongie, et al. Microsoft coco: common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[169] Maria Nazir, Sadia Shakil, and Khurram Khurshid. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. *Computerized medical imaging and graphics*, 91:101940, 2021.

[170] Scott W Atlas. *Magnetic resonance imaging of the brain and spine*, volume 1. Lippincott Williams & Wilkins, 2009.

[171] Ranjeet Kaur and Amit Doegar. Localization and classification of brain tumor using machine learning & deep learning techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(9):2278–3075, 2019.

[172] Ai central data science institute. `https://aicentral.acrdsi.org/`.

[173] Akshay S Chaudhari, Christopher M Sandino, Elizabeth K Cole, David B Larson, Garry E Gold, Shreyas S Vasanawala, Matthew P Lungren, Brian A Hargreaves, and Curtis P Langlotz. Prospective deployment of deep learning in mri: a framework for important considerations, challenges, and recommendations for best practices. *Journal of Magnetic Resonance Imaging*, 54(2):357–371, 2021.

[174] Akshay S Chaudhari, Zhongnan Fang, Feliks Kogan, Jeff Wood, Kathryn J Stevens, Eric K Gibbons, Jin Hyung Lee, Garry E Gold, and Brian A Hargreaves. Super-resolution musculoskeletal mri using deep learning. *Magnetic resonance in medicine*, 80(5):2139–2154, 2018.

[175] Ariel Benou, Ronel Veksler, Alon Friedman, and T Riklin Raviv. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced mri sequences. *Medical image analysis*, 42:145–159, 2017.

[176] David Y Zeng, Jamil Shaikh, Signy Holmes, Ryan L Brunsing, John M Pauly, Dwight G Nishimura, Shreyas S Vasanawala, and Joseph Y Cheng. Deep residual network for off-resonance artifact correction with application to pediatric body mra with 3d cones. *Magnetic resonance in medicine*, 82(4):1398–1411, 2019.

[177] Jens Kleesiek, Jan Nikolas Morshuis, Fabian Isensee, Katerina Deike-Hofmann, Daniel Paech, Philipp Kickingereder, Ullrich Köthe, Carsten Rother, Michael Forsting, Wolfgang Wick, et al. Can virtual contrast enhancement in brain mri replace gadolinium?: a feasibility study. *Investigative radiology*, 54(10):653–660, 2019.

[178] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[179] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for mr image reconstruction. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 647–658. Springer, 2017.

[180] N. Newell, JP Little, A. Christou, MA Adams, CJ Adam, and SD Masouros. Biomechanics of the human intervertebral disc: A review of testing techniques and results. *Journal of the Mechanical Behavior of Biomedical Materials*, 69:420–434, 2017.

[181] Chanyuan Liu, Jun Ran, John N Morelli, Bowen Hou, Yitong Li, and Xiaoming Li. Determinants of diurnal variation in lumbar intervertebral discs and paraspinal muscles: A prospective quantitative magnetic resonance imaging study. *European Journal of Radiology*, 160:110712, 2023.

[182] Merve Apaydin, Mehmethan Yumus, Ali Degirmenci, Serdar Kesikburun, and Omer Karal. Deep convolutional neural networks using u-net for automatic intervertebral disc segmentation in axial mri. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2022.

[183] Alessandro Liguori, Marco Pandolfi, Martina Gurgitano, Antonio Arrichiello, Letizia Di Meglio, Salvatore Alessio Angileri, Anna Maria Ierardi, Aldo Paolucci, Federica Galli, Elvira Stellato, et al. Image-guided percutaneous mechanical disc decompression for herniated discs: a technical note. *Acta Bio Medica: Atenei Parmensis*, 91(Suppl 10), 2020.

[184] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[185] Hua-Dong Zheng, Yue-Li Sun, De-Wei Kong, Meng-Chen Yin, Jiang Chen, Yong-Peng Lin, Xue-Feng Ma, Hong-Shen Wang, Guang-Jie Yuan, Min Yao, and et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from mri. *Nature Communications*, 13(1), 2022.

[186] Junyong Zhao, Liang Sun, Xin Zhou, Shuo Huang, Haipeng Si, and Daoqiang Zhang. Residual-atrous attention network for lumbosacral plexus segmentation with mr image. *Computerized Medical Imaging and Graphics*, 100:102109, 2022.

[187] Chuanpu Li, Tianbao Liu, Zeli Chen, Shumao Pang, Liming Zhong, Qianjin Feng, and Wei Yang. Spa-resunet: Strip pooling attention resunet for multi-class segmentation of vertebrae and intervertebral discs. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[188] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. Spineparsenet: Spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2021.

[189] Jhon Jairo Saenz-Gamboa, Maria De La Iglesia-Vayá, and Jon A. Gómez. Automatic semantic segmentation of structural elements related to the spinal cord in the lumbar region by using convolutional neural networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5214–5221, 2021.

[190] Pabitra Das, Chandrajit Pal, Amit Acharyya, Amlan Chakrabarti, and Saumyajit Basu. Deep neural network for automated simultaneous intervertebral disc (ivds) identification and segmentation of multi-modal mr images. *Computer Methods and Programs in Biomedicine*, 205:106074, 2021.

[191] Feng Zhao, Kai Dang, and Hanqiang Liu. Intervertebral discs localization and segmentation based on broad learning system and ipu-net. In *2021 3rd International Conference on Natural Language Processing (ICNLP)*, pages 248–254, 2021.

[192] Shumao Pang, Chunlan Pang, Zhihai Su, Liyan Lin, Lei Zhao, Yangfan Chen, Yujia Zhou, Hai Lu, and Qianjin Feng. Dgmsnet: Spine segmentation for mr image by a detection-guided mixed-supervised segmentation network. *Medical Image Analysis*, 75:102261, 2022.

[193] Xihe Kuang, Jason Pui Yin Cheung, Kwan-Yee K. Wong, Wai Yi Lam, Chak Hei Lam, Richard W. Choy, Christopher P. Cheng, Honghan Wu, Cao Yang, Kun Wang, Yang Li, and Teng Zhang. Spine-gflow: A hybrid learning framework for robust multi-tissue segmentation in lumbar mri without manual annotation. *Computerized Medical Imaging and Graphics*, 99:102091, 2022.

[194] Meiyan Huang, Shuoling Zhou, Xiumei Chen, Haoran Lai, and Qianjin Feng. Semi-supervised hybrid spine network for segmentation of spine mr images. *Computerized Medical Imaging and Graphics*, 107:102245, 2023.

[195] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[196] Karani Neerav, Chaitanya Krishna, Baumgartner Christian, and Konukoglu Ender. A lifelong learning approach to brain mr segmentation across scanners and protocols. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 476–484. Springer International Publishing, 2018.

[197] Frances Ryan, Karen López-Linares Román, Blanca Zufiria Gerbolés, Kristin May Rebescher, Maialen Stephens Txurio, Rodrigo Cilla Ugarte, María Jesús García González, and Ivan Macía Oliver. Unsupervised domain adaptation for the segmentation of breast tissue in mammography images. *Computer Methods and Programs in Biomedicine*, 211:106368, 2021.

[198] Jin Hong, Simon Chun-Ho Yu, and Weitian Chen. Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning. *Applied Soft Computing*, 121:108729, 2022.

[199] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[200] Alexander Bigalke, Lasse Hansen, and Mattias P Heinrich. Adapting the mean teacher for keypoint-based lung registration under geometric domain shifts. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 280–290. Springer, 2022.

[201] Daniel Franco-Barranco, Julio Pastor-Tronch, Aitor González-Marfil, Arrate Muñoz-Barrutia, and Ignacio Arganda-Carreras. Deep learning based domain adaptation for mitochondria segmentation on em volumes. *Computer Methods and Programs in Biomedicine*, 222:106949, 2022.

[202] Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang. Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures. *IEEE Transactions on Medical Imaging*, 42(1):245–256, 2023.

[203] Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S. Kevin Zhou. Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(3):633–646, 2023.

[204] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry L Prince, and Zongben Xu. Unsupervised mr-to-ct synthesis using structure-constrained cyclegan. *IEEE transactions on medical imaging*, 39(12):4249–4261, 2020.

[205] Hongfei Sun, Qianyi Xi, Rongbo Fan, Jiawei Sun, Kai Xie, Xinye Ni, and Jianhua Yang. Synthesis of pseudo-ct images from pelvic mri images based on an md-cyclegan model for radiotherapy. *Physics in Medicine & Biology*, 67(3):035006, 2022.

[206] Julián Alberto Palladino, Diego Fernandez Slezak, and Enzo Ferrante. Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images. In Jorge Brieva, Natasha Lepore, Marius George Linguraru, and Eduardo Romero Castro M.D., editors, *16th International Symposium on Medical Information Processing and Analysis*, volume 11583, page 1158302. International Society for Optics and Photonics, SPIE, 2020.

[207] Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince. Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical image analysis*, 72:102136, 2021.

[208] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

[209] Zhiming Cui, Changjian Li, Zhixu Du, Nenglun Chen, Guodong Wei, Runnan Chen, Lei Yang, Dinggang Shen, and Wenping Wang. Structure-driven unsupervised domain adaptation for cross-modality cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12):3604–3616, 2021.

[210] Yingying Xue, Shixiang Feng, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. Dual-task self-supervision for cross-modality domain adaptation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 408–417. Springer International Publishing, 2020.

[211] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3d multi-organ segmentation. In *Medical Image*

Computing and Computer Assisted Intervention – MICCAI 2020, pages 656–66. Springer International Publishing, 2020.

[212] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19786–19797, 2023.

[213] Donghyun Kim, Kuniaki Saito, Samarth Mishra, Stan Sclaroff, Kate Saenko, and Bryan A Plummer. Self-supervised visual attribute learning for fashion compatibility. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1057–1066, 2021.

[214] Ling Xiao and Toshihiko Yamasaki. Semi-supervised fashion compatibility prediction by color distortion prediction. *arXiv preprint arXiv:2212.14680*, 2022.

[215] Adrian Galdran, Katherine J. Hewitt, Narmin Ghaffari Laleh, Jakob N. Kather, Gustavo Carneiro, and Miguel A. González Ballester. Test time transform prediction for open set histopathological image recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 263–272. Springer Nature Switzerland, 2022.

[216] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2020.

[217] Chengwen Chu, Daniel L. Belavý, Gabriele Armbrecht, Martin Bansmann, Dieter Felsenberg, and Guoyan Zheng. Fully automatic localization and segmentation of 3d vertebral bodies from ct/mr images via a learning-based method. *PLOS ONE*, 10(11), 2015.

[218] Yasmina Al Khalil, Edoardo A Becherucci, Jan S Kirschke, Dimitrios C Karampinos, Marcel Breeuwer, Thomas Baum, and Nico Sollmann. Multi-scanner and multi-modal lumbar vertebral body and intervertebral disc segmentation database. *Scientific Data*, 9(1):97, 2022.

[219] Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. Medical image segmentation: A review of modern architectures. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 691–708. Springer, 2023.

[220] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[221] Michihito Nozawa, Hirokazu Ito, Yoshiko Ariji, Motoki Fukuda, Chinami Igarashi, Masako Nishiyama, Nobumi Ogi, Akitoshi Katsumata, Kaoru Kobayashi, and Eiichiro Ariji. Automatic segmentation of the temporomandibular joint disc on magnetic resonance images using a deep learning technique. *Dentomaxillofacial Radiology*, 51(1):20210185, 2022.

[222] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[223] Hao Du, Jiazheng Wang, Min Liu, Yaonan Wang, and Erik Meijering. Swinpa-net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[224] Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35:29582–29596, 2022.

[225] Paolo Zaffino, Sara Moccia, Elena De Momi, and Maria Francesca Spadea. A review on advances in intra-operative imaging for surgery and therapy: Imagining the operating room of the future. *Annals of Biomedical Engineering*, pages 1–21, 2020.

[226] M. Backhaus, G. Burmester, T. Gerber, W. Grassi, K. Machold, W. A. Swen, R. Wakefield, and B. Manger. Guidelines for musculoskeletal ultrasound in rheumatology. *Annals of the Rheumatic Diseases*, 60:641 – 649, 2001.

[227] YiRang Shin, Jaemoon Yang, Young Han Lee, and Sungjun Kim. Artificial intelligence in musculoskeletal ultrasound imaging. *Ultrasonography*, 40(1):30, 2021.

[228] Gianluca Smerilli, Edoardo Cipolletta, Gianmarco Sartini, Erica Moscioni, Mariachiara Di Cosmo, Maria Chiara Fiorentino, Sara Moccia, Emanuele Frontoni, Walter Grassi, and Emilio Filippucci. Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level. *Arthritis Research & Therapy*, 24(1):1–8, 2022.

[229] Mariachiara Di Cosmo, Maria Chiara Fiorentino, Francesca Pia Villani, Gianmarco Sartini, Gianluca Smerilli, Emilio Filippucci, Emanuele Frontoni, and Sara Moccia. Learning-based median nerve segmentation from ultrasound images for carpal tunnel syndrome evaluation. *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 2021.

[230] Philippe Burlina, Seth Billings, Neil Joshi, and Jemima Albayda. Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods. *PloS one*, 12(8):e0184059, 2017.

[231] Vincenzo Venerito, Orazio Angelini, Gerardo Cazzato, Giuseppe Lopalco, Eugenio Maiorano, Antonietta Cimmino, and Florenzo Iannone. A convolutional neural network with transfer learning for automatic discrimination between low and high-grade synovitis: a pilot study. *Internal and Emergency Medicine*, 16(6):1457–1465, 2021.

[232] Mariachiara Di Cosmo, Maria Chiara Fiorentino, Francesca Pia Villani, Emanuele Frontoni, Gianluca Smerilli, Emilio Filippucci, and Sara Moccia. A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet. *Medical & Biological Engineering & Computing*, 60(11):3255–3264, 2022.

[233] Yuichi Yoshii, Chunfeng Zhao, and Peter C. Amadio. Recent advances in ultrasound diagnosis of carpal tunnel syndrome. *Diagnostics.*, 10.(8.):596., 2020.

[234] Luca Padua, Daniele Coraci, Carmen Erra, Costanza Pazzaglia, Ilaria Paolasso, Claudia Loreti, Pietro Caliandro, and Lisa D Hobson-Webb. Carpal tunnel syndrome: Clinical features, diagnosis, and management. *The Lancet Neurology.*, 15.(12.):1273–1284., 2016.

[235] Gianluca Smerilli, Andrea Di Matteo, Edoardo Cipolletta, Sergio Carloni, Antonella Incorvaia, Marco Di Carlo, Walter Grassi, and Emilio Filippucci. Ultrasound assessment of carpal tunnel in rheumatoid arthritis and idiopathic carpal tunnel syndrome. *Clinical Rheumatology.*, 40.(3.):1085–1092., 2020.

[236] You-Wei Wang, Ruey-Feng Chang, Yi-Shiung Horng, and Chii-Jen Chen. MNT-DeepSL: Median nerve tracking from carpal tunnel ultrasound images with deep similarity learning and analysis on continuous wrist motions. *Computerized Medical Imaging and Graphics.*, 80.:101687., 2020.

[237] Muralikrishna Puttagunta and S Ravi. Medical image analysis based on Deep Learning approach. *Multimedia Tools and Applications.*, pages 1–34., 2021.

[238] Adel Hafiane, Pierre Vieyres, and Alain Delbos. Phase-based probabilistic active contour for nerve detection in ultrasound images for regional anesthesia. *Computers in Biology and Medicine.*, 52.:88–95., 2014.

[239] M. Alkhatib, A. Hafiane, Omar Tahri, P. Vieyres, and A. Delbos. Adaptive median binary patterns for fully automatic nerves tracking in ultrasound images. *Computer Methods and Programs in Biomedicine.*, 160.:129–140., 2018.

[240] O. Hadjerci, A. Hafiane, Donatello Conte, P. Makris, P. Vieyres, and A. Delbos. Computer-aided detection system for nerve identification using ultrasound images: A comparative study. *Informatics in Medicine Unlocked.*, 3.:29–43., 2016.

[241] Adel Hafiane, Pierre Vieyres, and Alain Delbos. Deep Learning with spatiotemporal consistency for nerve segmentation in ultrasound images. *arXiv preprint arXiv:1706.05870.*, 2017.

[242] Akhilesh Kakade and Jaysidh Dumbali. Identification of nerve in ultrasound images using U-Net architecture. In *2018 International Conference on Communication information and Computing Technology (ICCICT).*, pages 1–6., 2018.

[243] Ming-Huwi Horng, Cheng-Wei Yang, Yung-Nien Sun, and Tai-Hua Yang. DeepNerve: A new convolutional neural network for the localization and segmentation of the median nerve in ultrasound image sequences. *Ultrasound in Medicine & Biology.*, 46.(9.):2439–2452., 2020.

[244] Federico Perazzi, A. Khoreva, Rodrigo Benenson, B. Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, pages 3491–3500., 2017.

[245] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation.*, 9.:1735–1780., 1997.

[246] Raymond T Festen, Verena JMM Schrier, and Peter C Amadio. Automated segmentation of the median nerve in the carpal tunnel using U-Net. *Ultrasound in Medicine & Biology.*, 47.(7.):1964–1969., 2021.

[247] Chueh-Hung Wu, Wei-Ting Syu, Meng-Ting Lin, Cheng-Liang Yeh, Mathieu Boudier-Revéret, Ming-Yen Hsiao, and Po-Ling Kuo. Automated segmentation of median nerve in dynamic

sonography using deep learning: Evaluation of model performance. *Diagnostics*, 11(10):1893, 2021.

[248] Ingrid Möller, Iustina Janta, Marina Backhaus, Sarah Ohrndorf, David A Bong, Carlo Martinoli, Emilio Filippucci, Luca Maria Sconfienza, Lene Terslev, Nemanja Damjanov, et al. The 2017 EULAR standardised procedures for ultrasound imaging in rheumatology. *Annals of the Rheumatic Diseases.*, 76.(12.):1974–1979., 2017.

[249] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 39.:1137–1149., 2015.

[250] Oussama Hadjerci, Adel Hafiane, Donatello Conte, Pascal Makris, Pierre Vieyres, and Alain Delbos. Ultrasound median nerve localization by classification based on despeckle filtering and feature selection. In *2015 IEEE International Conference on Image Processing (ICIP).*, pages 4155–4159., 2015.

[251] M. Everingham, S. Eslami, L. Gool, Christopher K. I. Williams, J. Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision.*, 111.:98–136., 2014.

[252] Alessandro Casella, Sara Moccia, Dario Paladini, Emanuele Frontoni, Elena De Momi, and Leonardo S Mattos. A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation. *Medical Image Analysis.*, page 102008., 2021.

[253] Maria Chiara Fiorentino, Sara Moccia, Morris Capparuccini, Sara Giamberini, and Emanuele Frontoni. A regression framework to head-circumference delineation from US fetal images. *Computer Methods and Programs in Biomedicine.*, 198.:105771., 2021.

[254] Yinghao Zheng, Lina Qin, Taorong Qiu, Aiyun Zhou, Pan Xu, and Zhixin Xue. Automated detection and recognition of thyroid nodules in ultrasound images using Improve Cascade Mask R-CNN. *Multimedia Tools and Applications.*, pages 1–21., 2021.

[255] Laura J Brattain, Brian A Telfer, Manish Dhyani, Joseph R Grajo, and Anthony E Samir. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal radiology*, 43(4):786–799, 2018.

[256] Phillip M Cheng and Harshawn S Malhi. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of digital imaging*, 30(2):234–243, 2017.

[257] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[258] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.

[259] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.