Politecnico
di Bari

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

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

Final Dissertation

Explainable Deep Learning for Medical Image Processing: Computer-aided Diagnosis and Robot-assisted Surgery

by

Sardar Mehboob Hussain

Supervisor:
Prof. Vitoantonio Bevilacqua, Ph.D.
Co-supervisor:
Prof. Domenico Buongiorno, Ph.D.

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

Course n°35, 01/11/2019 - 31/01/2023

LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il sottoscritto **SARDAR MEHBOOB HUSSAIN** nato a **KOTLI AJK** il **02-02-1993**

residente a **Bari** in via **Della Resistenza 48D, 70125, BA** e-mail **sardarmehboob.hussain@poliba.it**

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

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

Explainable Deep Learning for Medical Image Processing: Computer-aided Diagnosis and Robot-assisted Surgery

DICHIARA

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

Luogo e data **Bari, 23-03-2023** Firma ______________________

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

CONCEDE

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

Luogo e data **Bari, 23-03-2023** Firma ______________________

Politecnico di Bari

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

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

Final Dissertation

Explainable Deep Learning for Medical Image Processing: Computer-aided Diagnosis and Robot-assisted Surgery

by

Sardar Mehboob Hussain

Referees:
Prof. Giovanni Dimauro
Prof. Leonardo Bocchi

Supervisor:
Prof. Vitoantonio Bevilacqua, Ph.D.

Co-supervisor:
Prof. Domenico Buongiorno, Ph.D.

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

Course n°35, 01/11/2019 - 31/01/2023

Dedicated to:
The one and only,
The Great,

"MY MOTHER"

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Vitoantonio Bevilacqua, and co-supervisor Domenico Buongiorno, for their invaluable guidance, support and encouragement throughout the course of this research. Their expertise and experience were instrumental in helping me to develop my research skills and shape my ideas into a cohesive project. I am deeply indebted for their patience, feedback, and commitment to my work.

I am thankful to all of my colleagues in our research group for creating a supportive and collaborative environment that facilitated my research.

I extend my heartfelt appreciation to every single person who has supported and encouraged me in any possible way during my journey. Your contributions have played a vital role, and I am forever beholden.

I am gratified to my niblings for their love, laughter, and joy. Their presence and tiny conversations have brought immense light and positivity into my life, and I am forever gladden for their innocence.

Finally and more importantly, I am grateful to my mother Zohra Fatima and my brother Matloob Hussain, for their endless love, support, and encouragement throughout my studies. Their unwavering belief and constant invigorate have made this journey possible.

Abstract

The recent advancements in the surging field of Deep Learning (DL) have revolutionized every sphere of life, and the healthcare domain is no exception. The enormous success of DL models, particularly with image data, has led to the development of several computer-aided diagnosis and clinical support systems. These intelligent imaging systems can help physicians in numerous medical tasks including classification and staging of the various diseases, image-guided surgical procedures, and many more. Additionally, the proliferation of medical datasets has further facilitated the applications of DL techniques in healthcare realm.

Moreover, all the perks DL offers are remarkable, however, DL architectures are typically blackbox, i.e. they hide the decision making mechanism, therefore, interpreting how the model arrived at a particular decision is hidden. Additionally, Convolutional Neural Networks (CNNs), which are most widely used DL techniques, are prone to adversarial examples, where small, imperceptible perturbations to the input data can cause the model to make incorrect predictions. These facts question the applicability of DL in healthcare sector where explainability holds paramount significance to build a trust on surging field of machine learning.

The concept of eXplainable Artificial Intelligence (XAI) brings forward the possibility of explaining the results of DL models and reveals how the models produce results. These techniques aim to improve the transparency and interpretability of AI models, which can enhance trust in their results and facilitate their adoption in clinical practice. XAI approaches have the potential to advance the understanding of complex medical image analysis tasks and improve the reliability of AI-based diagnosis and treatment planning.

The story does not end here, the XAI methods in the context of medical imaging generally produce saliency maps and compute feature importance to explain the results of DL models. The sensitive nature of healthcare industry, because of having the direct correlation with human life, questions the authenticity of XAI outcomes, and demands a qualitative and quantitative measure to evaluate these evaluation methods. Furthermore, heatmap visualizations

alone are often insufficient for achieving transparency and interpretability of DL models in medical imaging to foster the AI and biomedical synergy.

Inspired by the latest trends and contributions in light of the aforementioned concerns, this thesis designs, develops, and validates an interpretable and transparent intelligent clinical decision support system based on traditional machine and DL architectures, whose outcomes can be qualitatively and quantitatively explained with XAI methods. The thesis also comprises a segmentation and detection pipeline for image-driven surgical applications. These novel intelligent systems aims to assist the physicians and clinicians in image-guided diagnostic and treatment systems. The developed interpretable diagnostic frameworks offer wide range of applications and can be extended to several clinical scenarios.

Concerning the XAI, transparency and interpretability of CNN architectures are achieved through two families of XAI methods, i.e. perceptive and mathematical XAI. Furthermore, within each of these XAI families, two explanation frameworks are employed. These methods facilitated to investigate the reliability of features and learning process, to critically analyse various CNN architectures and XAI methods, and to compare the outcomes of both XAI pipelines.

To further highlight the applications of DL in the image-guided surgical domain, a case study has been performed on image-guided surgical procedures and interventions. The case study also encompasses a detailed investigative study of public datasets and presents the legal and ethical issues of DL-driven image-guided surgery. The study additionally underlines the risks and limitations towards the autonomous systems and provides the future perspective.

Finally, the second case study investigates the qualitative and quantitative evaluation of the XAI techniques in regards to the medical images. The case study also sheds light on the evaluation measures, metrics for XAI, quality of explanation, types of explanation, and few more.

The clinical efficacy of the developed solutions is evaluated through comparison with existing state-of-the-art methods, and is further validated through consultation with physicians where feasible. The datasets incorporated during the study are either obtained from the online open source platforms or collected from local health institutions.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Deep learning (DL) has undergone significant development in recent years, leading to its widespread application across various sectors, including healthcare [1, 2]. These advancements brought a transformative impact in medical realm, resulting in numerous innovations and improvements. The DL has emerged as a promising computational approach for the automatic detection, classification, and segmentation of various diseases thorough the analysis of diagnostic medical images, thus enabling the Computer-aided Diagnosis (CAD), clinical decision support systems, and surgical robotics among several others [3–6]. The DL methods along with the traditional image processing techniques have already been established as an effective approach to automatically analyze medical images for diagnosis and monitoring [7–10]. Additionally, the contemporary availability of the image datasets has boosted the interdisciplinary synergies of biomedical engineers and physicians in healthcare industry.

Moreover, before the advent of the modern image modality capturing systems, the physicians and the surgeons mostly relied upon simple cameras and naked eyes to study the internal behaviour of the organs. Today, the most common imaging modalities include X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), and Positron Emission Tomography (PET) [11, 12].

However, even the modern imaging modalities required intensive preprocessing and feature engineering [13]. Thanks to the DL, this laborious, time consuming, and cost intensive task is no more as tedious as heretofore.

Additionally, the basic underlying principle of the DL mimics the (functionality of) biological neuron, that connects with a complex layered structure, learns from generalization, and keeps the neuron-associated weights updated. One of the most powerful models of the

DL is believed to be the Convolutional Neural Network (CNN). The introduction of the CNN can be traced back to early 1960s [14], which has led to the development of several highly efficient diagnostic systems [15].

Furthermore, the subsequent rise of DL has also assisted the surgeons in the operating room in several different ways [16]. This successful incorporation has paved the way for Robot Assisted Surgery (RAS) and other surgical planning systems [17]. The purpose of RAS, as the name suggests, is not to replace the surgeons and physicians but to assist them in order to achieve higher proficiency in security and safety of the undergoing patients in preoperative, intraoperative, and postoperative surgical procedures [18, 19].

Image driven DL methods for robotic surgery have already taken care of the instrument detection and segmentation [20, 21], gesture recognition [22], workflow analysis [23], skill assessment [24], and many more [25–28] to facilitate the semi-autonomous RAS.

Moreover, the development of a fully autonomous image-guided surgical system, where the direct involvement of the surgeon is seldom required, is foreseeable task for the DL models. The DL has ultimately proven the enormous success in Minimally Invasive Surgery (MIS) systems. The very first RAS system i.e. da Vinci surgical system, introduced in the year 2000, has successfully performed around $1,594,000$ surgical procedures in 2021 [29] with an increase of 28% from the previous year ($1,243,000$ in 2020) and is expected to perform $12-15\%$ more in the following part of the year. The MIS reduces the post-surgery trauma, minimises the hospital stay, improves recovery, and avoids potential risk of contagion [30].

In spite of the enormous success in all of the aforementioned fields, the complex nature of the DL techniques hides any possible information of the underlying decision mechanism [31, 32], which questions its application in the healthcare domain where explainability holds paramount significance to build a trust on decisions made by inevitably booming Artificial Intelligence (AI).

The super successful DL models come with blackbox nature. The eXplainable Artificial Intelligence (XAI) brings forward the possibility of explaining the results of the blackbox DL models and reveals how the models produce results. Generally, XAI is supposed to fit a model onto four basic attributes [33]:

- *Transparent*: open to the degree where humans can understand the decision-making mechanism.
- *Justifiable*: the decision can be supported or justified along each step.
- *Informative*: to provide reasoning and allow reasoning.

- *Uncertainty yielding*: does not follow hard-coded structure, but open to change.

XAI has drawn a tremendous amount of attention in the recent past and to comprehend the importance of such methodologies in the clinical field, where AI is spreading fast [34], has become indispensable. The symbiosis of AI and XAI is extremely fascinating yet challenging, because, as it can be easily envisaged, a more complex AI model that can reach high-level performance is less interpretable than, for example, a simple rule-based model, however, at the cost of unsatisfactory outcomes.

The interpretability and explainability have largely been studied and categorised into two families of methods, namely, perceptive interpretability and mathematical interpretability [34]. The perceptive XAI is responsible for bringing a straightforward visualisation of the top contributing features that affect the final predictions, whereas the mathematical interpretability provides insights into the used models and portrays the features that are employed to make the final predictions. The former is used to study the feature-level classification behavior (the importance of a particular region towards classification) of the DL architectures, whereas the latter is used to study the clustering capabilities of the DL networks.

More importantly, as the definition of XAI states, the purpose of XAI is to make the DL decisions understandable to human. Merely relying upon saliency maps and feature contribution values lack the actual definition of XAI, particularly in medical imaging domain. The debate to make the XAI decision understandable to an expert or to a common human can be considered progressive, however, the requirement to explain the outcomes and decision mechanism of DL architectures remains intact in either case.

Nevertheless, unlike other domains, the medical domain can not merely depend upon machine trust, technology reliance, mutual understanding, and argumentation about the XAI methods. In literature, several methods have been proposed for the evaluation and quantification of XAI methods, however, there is no one compact and generalised method for quantitatively evaluating the XAI methods on different types of medical images. A common practice has been seeking help from the clinicians to evaluate the explanations generated by XAI methods, however, this method is prone to errors, time consuming, labour intensive, and experience demanding. Nevertheless, the visualisation of top contributing features, spotlighting the important regions, and computing numerous scores of contribution towards decision have long been discussed, alongside what is required is a quantitative and qualitative method to measure the effectiveness of an explanation.

Moreover, one of the most proficient and prudent questions is to define what is a good explanation. What defines/declares and makes an explanation good is another relevant and interesting question to raise.

Additionally, the explanation of explainable methods is also context dependent that arises another question, i.e. an explanation must be understandable and interpretable but to whom? To general public? To experts? To machines? Or to whom? An interpretation of the cancer classification model on breast images is only understandable to physicians and relevant experts. The evaluation of the XAI methods depends upon the end user of the application and the sufficiency of the quality of explanation depends upon the application area, explanation purpose, and the targeted audience. Therefore, all these questions open new horizons and direct to the context dependent applications.

1.2 Motivation

The revolutionary advent of DL has technologically redefined the working principles of all spheres of human life. The healthcare domain has also seen marvelous progress in the recent decade, particularly after the introduction of iconic work by Krizhevsky et al. [35]. On top of this, the large scale availability of medical imaging data has further boosted the development of CAD systems.

However, in spite of all the advancements, there still exists plenty of room for further improvements and innovations in DL applications in CAD, image-guided surgery, and other autonomous systems for the scientific community. The classification, segmentation, and identification of various diseases on medical image data have not reached to the full potential. Apart from the intrinsic bias in the data collection procedures and protocols, which pose great threat to medical domain, and besides the benign vs malignant cancer classification, there are several cancer types and stages which additionally vary with respect to shape, size, and other morphological patterns.

Additionally, the blur images and videos generated by camera are often misinterpreted and mislabelled by physicians and AI systems, because of the presence of smoke, shade of tools, shapes of lesion, plasma stains, vessels, and many more [36–39].

The breast cancer is morphologically categorized into several varying shapes based on cancer growth pattern, named as round, oval, lobulated, irregular, and architectural distortion [40, 41]. The availability of large scale data for each independent morphological category is cumbersome, which invites the option of artificial data generation or incorporation of pretrained networks.

In spite of the enormous success, the blackbox nature of the DL techniques hides any possible information of the underlying decision mechanism [31, 32], which questions its usage in the healthcare domain where explainability holds paramount significance to build a trust on decisions made by surging AI.

The General Data Protection Regulation (GDPR) by European Union states the concise and transparent information provision and privacy protection of users [42]. The clause 13 and 14 empower the users (i.e. patients) to ask for decision making mechanism and other relevant information. XAI brings forward the possibility of explaining the results of DL models and reveals how the models produce these highly accurate results.

Additionally, as the definition of XAI states, the purpose of XAI is to make the DL decisions understandable to human. Merely relying upon saliency maps and feature contribution values lack the actual definition. The debate to make the XAI decision understandable to an expert or to a common human can be considered progressive, however, the requirement to explain remains intact in either case.

Moreover, unlike other domains, the medical imaging domain can not rely on machine sense, mutual understanding, and argumentation about the XAI methods, therefore, evaluating the effectiveness of the XAI techniques is indispensable. In literature, several methods have been proposed for the evaluation and quantification of XAI methods, however, there is no one compact and generalised method for quantitatively evaluating the XAI methods on different types of medical images. A common practice has been seeking help from the clinicians to evaluate the explanations generated by XAI methods, however, this method is prone to errors, time consuming, labour intensive, and experience demanding. Nevertheless, the visualisation of top contributing features, spotlighting the important regions, and computing numerous scores of contribution towards decision have long been discussed, alongside what is required is a quantitative and qualitative method to measure the effectiveness of an explanation.

In light of all the aforementioned issues, the concise motivation of this thesis lies in the conceptualisation, design, development, and validation of an intelligent system that can be fed with medical images to support the clinical decision systems. The development of transparent systems under the explainability methods in order to create trust for AI in medical realm and to provide maximum assistance to the physicians. Finally, a qualitative and quantitative evaluation of the results produced by XAI methods on the decision mechanism of DL architectures, so that the explanation and interpretation can be validated in the context of medical imaging domain.

In this regard, several intelligent applications have been designed, developed, and validated in this thesis work. The XAI evaluations are additionally presented for the developed applications to validate the outcomes. Finally, two independent case studies for quantitative and qualitative evaluations of XAI and applications of image based DL models in robotic surgery are presented. The datasets employed in this study are either provided by collaborators or acquired from public repositories.

1.3 Contributions

In light of the above discussed issues, the main objective of the thesis was to design, develop, and validate an interpretable and transparent intelligent clinical decision support system based on DL architectures, whose outcomes can be explained with XAI methods. The novel intelligent systems were aimed to assist the medical experts and physicians in the CAD systems and surgical procedures. Such intelligent systems have been designed, developed, and validated with the novel DL techniques and the results are further interpreted with several XAI models. The developed interpretable diagnostic frameworks offer wide range of applications and can be extended to several clinical scenarios. The devised intelligent systems are compared with the state-of-the-art approaches already discussed in the literature. The applicability of the proposed solutions has also been validated with the help of physicians and the domain experts where required.

The technical contributions of this study are threefold, each of which are further subdivided into several smaller chunks. The initial part of primary contribution includes the development and validation of a CNN-based DL framework for the classification of breast lesions according to the shape by analyzing the related Region of Interest (RoI) on DBT images. Considering the shapes of cancerous masses, the Breast Imaging Reporting & Data System (BIRADS) classification of the American College of Radiology, which is the most commonly employed in the clinical and digital breast tomosynthesis settings, has been considered [43]. Similarly, concerning the surgical procedure part, a framework to address the tasks of vertebrae segmentation and identification by exploiting both DL and classical machine learning methodologies is also proposed. The presented solution comprises two phases: a binary fully automated segmentation of the whole spine, which exploits a 3D CNN, and a semi-automated procedure that allows locating vertebrae centroids using traditional machine learning algorithms. Likewise, a novel optimization formulation for automatic contour delineation of the prostate gland from Transrectal Ultrasound (TRUS) images, to find the best superellipse a deformable model, that can accurately represent the prostate shape, is devised. The advantage of the proposed approach is that it does not require extensive annotations, and can be used independent of the specific transducer employed during prostate biopsies.

Furthermore, the second major part of thesis contributions, which has been given a considerable attention, is the incorporation of interpretability and explanability of the CNN architectures using two families of XAI methods. This includes investigation on the applicability of both perceptive and mathematical XAI methods; investigation on the reliability of features and learning processes and correlation with the overall DL model performance; a comprehensive comparison of the CNN architectures and the XAI methods in order to guide the engineers and the radiologists interested in implementing DL-driven CAD systems; and an exhaustive comparison of outcomes of XAI with several different methods.

Moreover, the thesis also comprises two independent case studies to further support the applications of DL in the medical imaging domain. The former case study has been performed on image-guided surgical procedures and interventions. The case study also encompasses a detailed investigative study of public datasets and presents the legal and ethical issues of image-driven RAS, and further highlights the risks and limitations towards the autonomous systems.

Finally, considering the sensitive nature of healthcare domain, XAI presents visual and textual explanations on the outcomes of DL methods applied on medical images. However, merely visualising the top contributing features and highlighting the important regions on images seldom make a DL model interpretable. The requirement of a qualitative and quantitative metric to evaluate the explanation of XAI methods is indispensable. In the second case study, the qualitative and quantitative evaluation of the XAI techniques has been studied and investigated in regards to the medical images. The case study also sheds light on the evaluation measures, metrics for XAI, quality of explanation, types of explanation, and few more.

These contributions resulted in shape based breast cancer classification framework [44], vertebrae segmentation and identification [45], prostate segmentation and registration [46], explainability of CNN models on breast morphological classification [44], DL driven image-guided surgery [47], and evaluation of XAI outcomes of DL architectures on medical images.

1.4 Thesis Structure

This thesis is presented into two major parts comprising nine chapters in total. The Part I deals with the applications of RAS and image-guided surgical systems, whereas, the Part II presents the CAD and clinical decision support systems along with the XAI and evaluation of XAI. Starting from the introductory background of the domain, motivation towards the study, and the technical contributions inscribed in the Chapter 1. The second chapter comprises the state-of-the-art in the applicability of DL in medical imaging domain with a particular emphasis on CNNs and performance measuring metrics 2.

The Part I starts by illustrating the Chapter 3, that spans the study on the applications of DL in the image-guided surgery, followed by the legal, ethical, and technological challenges towards the autonomous systems. The chapter also discusses the widely applied datasets in image-guided interventions and the limitations of the existing autonomous systems. Chapter 4 and Chapter 5 present the image-guided surgical applications, where the former inscribes the prostate segmentation and identification, and the latter contains vertebrae segmentation based upon traditional machine learning and DL models.

The Part II embraces the CAD systems along with the XAI and evaluation of XAI methods. Chapter 6 presents the devised morphological classification frameworks for breast cancer morphology. The chapter also highlights the first set of contributions of this thesis work by comprehensively explaining the induced methodologies, employed frameworks, achieved results, limitations, and discussion. Furthermore, the explainability and interpretability of the DL models applied on the devised workflows along with the outcomes are provided in the Chapter 7. The chapter sheds light on both, mathematical and perceptive XAI methods within the realm. The evaluation protocols of XAI methods, the recent contributions, and the way forward are described in the Chapter 8 with a precise focus on interpretability of explainable DL models in medical imaging domain.

Finally, Chapter 9 concludes with the final remarks and highlights the prominent findings and offers a future perspective of the study for potential research community.

Chapter 2

State-of-the-art

2.1 Deep Learning: A Broader Picture

With the advancement in technology and the increasing amount of medical images being generated, DL has become an essential tool for automating medical tasks such as image segmentation, diagnosis, and detection of diseases [9, 10]. DL has revolutionized the healthcare realm by enabling faster, more accurate, and cost-effective diagnoses, and ultimately improving patient outcomes [48].

Applications of DL in medical imaging domain range from automated segmentation of structures, diagnosis and classification of diseases, detection of tumors, abnormalities, lesions, and many more [49–52] to support the CAD, RAS, image-guided surgery, and other intelligent imaging systems [47, 53]. Most common DL methods for medical image analysis include: CNNs, which are particularly well suited for image analysis tasks, Recurrent Neural Networks (RNNs), which are useful for processing sequential data such as time-series medical images, and Generative Adversarial Networks (GANs), which are also getting popularity in medical imaging, as they can generate new images based on the training dataset, which can be used to augment the training dataset, and improve the model performance [54–56].

Overall, deep learning is showing great potential in medical image analysis and is expected to have a significant impact on the field of medical imaging and healthcare in general.

The section below encircles the fundamental concepts of DL in the context of medical image analysis and inscribes the formal introduction of the methods and techniques that appear in the relevant literature of the thesis scope.

2.1.1 Learning Paradigms

DL is a subset of machine learning that is inspired by the structure and function of the human brain, specifically, artificial neural networks. There are several different learning paradigms used in DL, each with their own advantages and disadvantages. These paradigms include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [57, 58].

2.1.1.1 Supervised Learning

Supervised learning is the most common and well-understood paradigm in DL. As the name suggests, supervised learning algorithms are fed with complete information under the managed supervision and the model makes decision based on the inputs. A supervised model is given input data along with the relevant labels and it learns by finding the relevant patterns in the data [59–61]. It involves training a model on labeled data to make predictions about new, unseen data. The process of supervised learning can be broken down into several steps:

- **Collect and prepare a labeled dataset:** This dataset consists of input-output pairs, where the input is typically a feature vector, and the output is the corresponding label.

- **Define a model architecture:** This involves defining the structure of neural network to make predictions. Common architectures include feedforward neural networks, CNN, and RNN.

- **Train the model:** Then comes the training of the model over the labeled dataset using an optimization algorithm, such as stochastic gradient descent. The goal here is to minimize the difference between the predicted labels and the ground-truth labels.

- **Evaluate the model:** The performance of the devised model is evaluated on a separate but labeled test dataset. Most common metrics used to evaluate the performance of a supervised learning model include accuracy, precision, recall, and F1-score.

- **Make predictions:** Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. This is achieved by forwarding the input data through the trained model and interpreting the output.

Supervised learning is used in a variety of applications, including image classification, natural language processing, and speech recognition. It is also the foundation for other learning paradigms, such as semi-supervised learning and reinforcement learning, described hereunder.

2.1.1.2 Unsupervised Learning

Unsupervised learning is a learning paradigm in deep learning that involves training a model on unlabeled data to discover patterns or features in the data. Unlike supervised learning, unsupervised learning does not have explicit output labels, and the model is not given any guidance on what the correct output should be. Instead, the model learns to extract useful information from the input data on its own [59–61].

The process of unsupervised learning can be broken down into several steps:

- **Collect and prepare an unlabeled dataset:** The dataset here consists of input data only, without any corresponding labels or any other information.

- **Define a model architecture:** The architecture of the network employed in unsupervised learning is typically simpler than that used in supervised learning, as it does not need to make predictions. Common architectures include autoencoders, generative models, and clustering algorithms.

- **Train the model:** The model is trained on the unlabeled dataset using an optimization algorithm, such as stochastic gradient descent. The goal is to discover important patterns and features in the data.

- **Evaluate the model:** The performance of the model is evaluated based on the quality of the patterns or features it has discovered. Common metrics used to evaluate the performance of an unsupervised learning model include reconstruction error, log-likelihood, clustering accuracy, among many others.

- **Use the model:** Once the model is trained, it can be used for tasks such as data compression, anomaly detection, and data generation.

  Unsupervised learning is used in a variety of applications, including dimensionality reduction, anomaly detection, and feature learning. It can also be used in conjunction with other learning paradigms, such as supervised learning and semi-supervised learning, to improve the performance of a model.

2.1.1.3 Semi-supervised Learning

Semi-supervised learning combines the benefits of both supervised and unsupervised learning paradigms. It involves training a model on a small amount of labeled data and a large amount of unlabeled data. The idea behind this is that the model can leverage the information contained in the large amount of unlabeled data to improve its performance on the labeled data [60, 61].

The process of semi-supervised learning can be broken down into several steps:

- **Collect and prepare a dataset:** The dataset consists of a small amount of labeled data and a large amount of unlabeled data.
- **Define a model architecture:** The architecture of the network used in semi-supervised learning can be the same as that used in supervised learning, or it can be a more complex architecture that can take advantage of the additional unlabeled data.
- **Pre-train the model:** The model is first pre-trained on the large amount of unlabeled data using an unsupervised learning algorithm, such as an autoencoder or a generative model.
- **Fine-tune the model:** Once the model is pre-trained, it can be fine-tuned on the small amount of labeled data using a supervised learning algorithm.
- **Evaluate the model:** The performance of the model is evaluated on a separate test dataset, which is also labeled. Common metrics used to evaluate the performance of a semi-supervised learning model include accuracy, precision, recall, and F1-score.
- **Make predictions:** Once the model is trained and evaluated, it can be used to make predictions on new, unseen data.

Semi-supervised learning is particularly useful when labeled data is scarce or expensive to obtain. It can also be used to improve the performance of a model trained on a small amount of labeled data, by leveraging the information contained in a large amount of unlabeled data.

2.1.1.4 Reinforcement Learning

Reinforcement learning is a DL paradigm where an agent learns to make decisions by interacting with its environment and receiving feedback in the form of rewards or penalties. The agent's goal is to learn a policy, which is a mapping from states of the environment to actions, that maximizes the cumulative reward over time [60, 61].

The process of reinforcement learning can be broken down into several steps:

- **Define the environment:** The environment consists of states, actions, and a reward function. The states represent the current situation of the agent, the actions are the choices an agent can make, and the reward function provides a scalar feedback signal indicating how good or bad the agent's actions are.

- **Define the agent:** The agent consists of a policy, which is a mapping from states to actions, and a value function, which estimates the expected cumulative reward of a policy. The agent also has a memory or history of its past interactions with the environment.
- **Interact with the environment:** The agent starts in an initial state, and at each step, it selects an action according to its current policy and receives a scalar reward from the environment. The agent then transitions to a new state, and this process continues.
- **Update the agent's policy:** The agent uses the rewards it receives to update the policy and value function. This process is typically done using a variant of gradient descent called Q-learning or policy gradient methods.
- **Evaluate the agent:** The performance of the agent is evaluated by measuring the cumulative reward it receives over time, or by measuring the asymptotic performance of the learned policy.

Reinforcement learning is used in a variety of applications, including robotics, game playing, and decision making. Reinforcement learning is also used in fields like finance, transportation and logistics, healthcare and manufacturing. Reinforcement learning is particularly useful when the environment is stochastic or non-stationary, and the agent must adapt to changing conditions.

2.1.2 Artificial Neural Network

Most of the DL models are based on neural network architecture that is verily inspired by the complex structure of human brain. At its core, a neural network is a mathematical model that is designed to recognize patterns in data. The basic building block of a neural network is the neuron, which is a simple processing unit that takes in inputs, performs a computation on them, and produces an output. Neurons are connected to one another in layers, and the output of one layer is fed as input to the next layer. This forms a network of neurons, which collectively can perform complex computations. A typical architecture of an artificial neural network is provided in the Figure 2.1.

There are different types of neural networks, such as feedforward neural networks and recurrent neural networks. Feedforward neural networks have a simple structure where

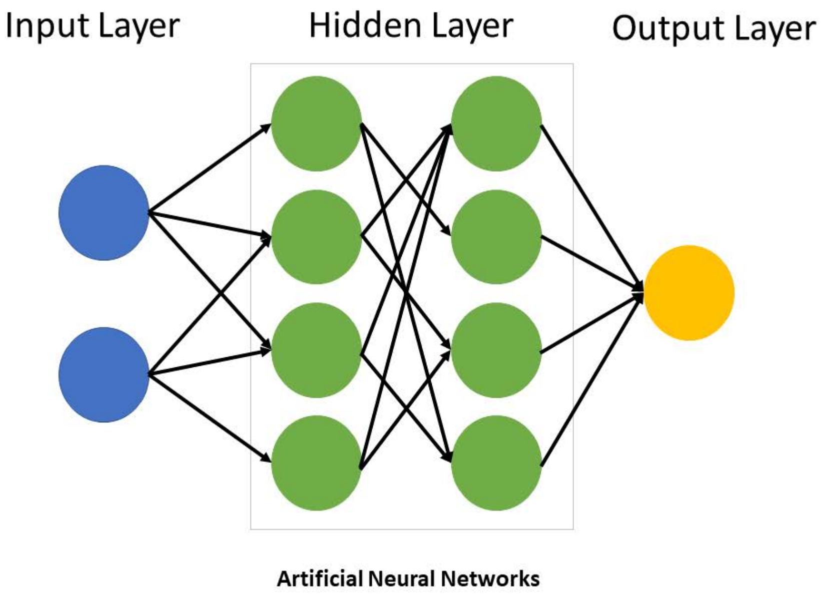


Fig. 2.1 Typical architecture of an artificial neural network

information flows in one direction from input layer to output layer. In contrast, recurrent neural networks have a looped structure where the output of a neuron is fed back into itself, allowing the network to retain information from previous time steps.

The process of training a neural network involves adjusting the parameters of the network, such as the weights on the connections between neurons, so that it can accurately perform a given task. This is done by presenting the network with a set of input-output pairs, called the training set, and adjusting the weights to minimize the difference between the network's output and the desired output. Once the network is trained, it can be used to make predictions on new, unseen data.

The DL has revolutionized the traditional machine learning by illuminating the manual feature extraction process. Furthermore, the introduction of the back propagation algorithm has enabled researchers to compute the impact each parameter imposes on the objective function [62]. The back propagation has further enriched the neural networks and made the computation faster, easier and better.

Until the introduction of layer over layer training of deep neural network, the training of the neural network was widely believed to be quite tedious and ineffective. However, Bengio et al. [63] proposed a mixed of unsupervised training during layer over layer training and supervised training while fine tuning at two different stages which showed considerable results. The algorithms trained in this manner include autoencoders and belief networks, which are still considered complex because of hectic process to reach substantial results.

2.1.3 Recurrent Neural Network

The neural network models appear to have another class known as RNN models. An RNN follows the sequential feeding of the input data. The RNN models are great improvement for time series sequential input data problems. The internal state in the RNN model, also called the memory of the neuron, saves the leading information coming from the previous computations.

The implementation of the RNN in the image driven computer-assisted methods has not been much appreciated in the literature, however, its successful adoption in the natural language processing tasks makes it standout. Another worth noting point of the RNN is its ability to work on variable length input data. The use of the RNN in the robotic systems driven by images and the kinematic data has been increasing over the time.

2.1.4 Convolutional Neural Networks

The most widely used DL model is CNN which has proven its applicability in image processing applications [15, 20, 64]. The generally accepted common CNNs based models include VGG16-19, ResNet, Inception, Xception, MobileNet, EfficientNet and many more. The key differences between a Multilayer Perceptron (MLP) and a CNN model are the inclusion of pooling layer in the CNN, sparsely connected layers instead of fully connected, and small associated weights of the layers that particularly help in dealing with image data.

The main building blocks of the CNN are convolutional layer, pooling layer, normalization layer, dense layer, dropout layer and activation layer, which along with their nature and responsibilities are described in the Table 2.1. The optimal number of layers in a network depends upon the nature of the problem that a network has to deal with, however, in general, there is no fixed number of layers, and hence, it is a matter of search to figure out the optimal number given a certain problem. To avoid the possible trade-off between the computational complexity and the performance, different numbers of the layers and neurons can be considered over repetitive iterations.

The networks with fewer number of layers and trainable parameters take less time, however, at the expense of lower accuracy. These type of models may not reach to full potential by modeling all the required parameters. On the other side, an overly populated network will provide better accuracy results but can also learn unnecessary features which will result in overfitting of the network. This type of model will perform poor on unforeseen data. The solution of the above problems is provided by the pretrained networks [65].

2.1.5 Major Building Blocks

The major building blocks of a DL architecture are presented hereunder.

2.1.5.1 Layer

As the name suggests, the principal operation of CNN is convolution, however, a number of additional layers are added namely dense, dropout, pooling layers and few more to the model to improve performance. The types of layers, the relevant hyperparameters, and their work is summarized in the Table 2.1.

2.1.5.2 Cost Function

The cost function or the loss function describes how well a model has performed with respect to the ground truth. A number of loss functions have been used in literature depending upon the operations to be performed by model over a specific data. The cross-entropy is most widely used loss function in classification problems [66].

2.1.5.3 Performance Measuring Metrics

The performance measuring metrics are important part of CNN which are the measuring scales that quantify the performance of the model.

The most commonly used metrics for the classification task include accuracy, precision, recall, and f1-score. The segmentation and the object detection tasks may have additional measuring parameters depending upon the nature and the definition of the problem.

2.1.6 Common Frameworks and Libraries

There exist numerous DL frameworks that facilitate to design, train, and validate neural networks using several interfaces. These frameworks and libraries include but not limited to TensorFlow, PyTorch, MATLAB, NVIDIA Caffe, Chainer, Theano, and Keras.

These high level interfaces help researchers, mathematicians, scientists, and developers to implement the complex architectures of deep neural networks to solve the various real world problems. Few of the most commonly used frameworks and libraries are listed below.

2.1.6.1 TensorFlow

An end-to-end open source library developed by Google Brain team, supports the numerical computation and analysis, is extensively used library that works with both CPU and GPU. The programming interface of TensorFlow is limited to Python and C++ [67].

2.1.6.2 PyTorch

PyTorch has received a tremendous amount of attention by the researchers and the developers because of its ability of easily implement the complex architectures of DL models [68]. Additionally, it also supports the tensor manipulations, e.g. NumPy computations.

2.1.6.3 MATLAB

MATLAB is well-known mathematical framework which is highly regarded in scientific society. It offers great visualization tools and is not limited to DL and neural networks [69]. The high-level features in the MATLAB do not require high level of expertise to implement. The CUDA code is automatically generated by MATLAB from simple code.

2.1.6.4 NVIDIA Caffe

It largely supports the GPU based computations. NVIDIA Caffe is worthy contribution of Berkeley Vision and Learning Center to the developer community. The main aims behind the Caffe development were speed and modularity [70].

2.1.6.5 Keras

Keras is another product by Google engineers which is deemed fruitful for beginners. The four main basic principles were considered during the development of Keras including modularity, minimalism, extensibility, and Python based [71].

2.2 Deep Learning Models for Classification

2.2.1 Convolutional Neural Networks

CNN are a type of DL neural network that are specifically designed for image classification tasks. A typical CNN is composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers, which along with their nature and responsibilities are described in the Table 2.1. An architecture of the CNN is provided in the 2.2. The optimal number of layers in a network depends upon the nature of the problem that a network has to deal with, however, in general, there is no fixed number of layers, and hence, it is a matter of search to figure out the optimal number given a certain problem. To avoid the possible trade-off between the computational complexity and the performance, different numbers of the layers and neurons can be considered over repetitive iterations.

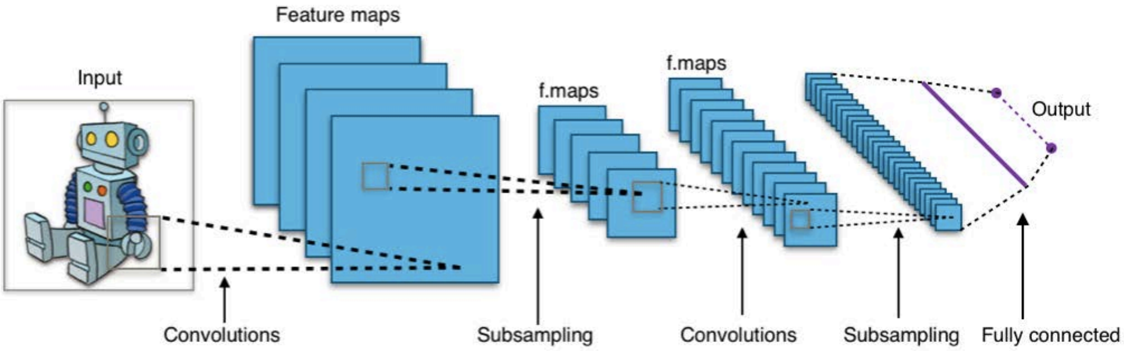


Fig. 2.2 A typical architecture of a convolutional neural network

- **Convolutional layers:** The primary building block of CNNs, these layers are responsible for extracting features from the input image. Convolutional layers use a set of filters (also called kernels or weights) that are convolved with the input image to produce feature maps. The filters are learned during the training process.

- **Pooling layers:** These layers are used to reduce the spatial resolution of the feature maps produced by the convolutional layers. This reduces the computational complexity of the network, and also makes it more robust to small translations and deformations in the input image. Two types of pooling are widely used: max pooling and average pooling.

- **Fully connected layers:** These layers are used to classify the image based on the features extracted by the convolutional and pooling layers. They are composed of multiple neurons (also called units) that are connected to all the neurons in the previous layer. The output of these layers is a set of scores for each class in the classification task.

The CNN has proven its applicability in image processing applications [15, 20, 64]. The generally accepted common CNNs based models include VGG16-19, ResNet, Inception, Xception, MobileNet, EfficientNet and many more. The key differences between a MLP

Table 2.1 The major building blocks of the CNN model along with the nature and the responsibilities

	Hyperparameter	Type of Layer	Responsibility
Convolutional Layers	Kernel size Stride Padding	1D Convolutional 2D Convolutional 3D Convolutional	Feature extraction
Pooling Layer	Pool size Padding	Max pooling Average pooling Global average pooling	Feature extraction Dimension reduction
Normalization Layer	Momentum Epsilon Beta	Batch normalization Instance normalization Group normalization Layer normalization	Input standardization
Dense Layer	Number of nodes	-	Fully connected layer
Dropout Layer	Rate	-	Overfitting avoidance
Activation Layer	Activation Function	ReLU Sigmoid Softmax	Activation function

and a CNN model are the inclusion of pooling layer in the CNN, sparsely connected layers instead of fully connected, and small associated weights of the layers that particularly help in dealing with image data.

The networks with fewer number of layers and trainable parameters take less time, however, at the expense of lower accuracy. These type of models may not reach to full potential by modeling all the required parameters. On the other side, an overly populated network will provide better accuracy results but can also learn unnecessary features which will result in overfitting of the network. This type of model will perform poor on unforeseen data. The solution of the above problems is provided by the pretrained networks [65]. Few of the most widely applied DL networks in the context of medical imaging are described hereunder.

2.2.1.1 AlexNet

AlexNet is a CNN that was developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton in 2012 [35]. It was the first CNN to achieve state-of-the-art results on the ImageNet image classification dataset. AlexNet is composed of 8 layers: 5 convolutional layers, 2 fully connected layers and a final softmax layer. It uses ReLU as activation function and dropout as regularization technique. AlexNet is considered to be the pioneer of DL in computer vision and is still widely used today. An example architecture of the network is presented in the Figure 2.3.

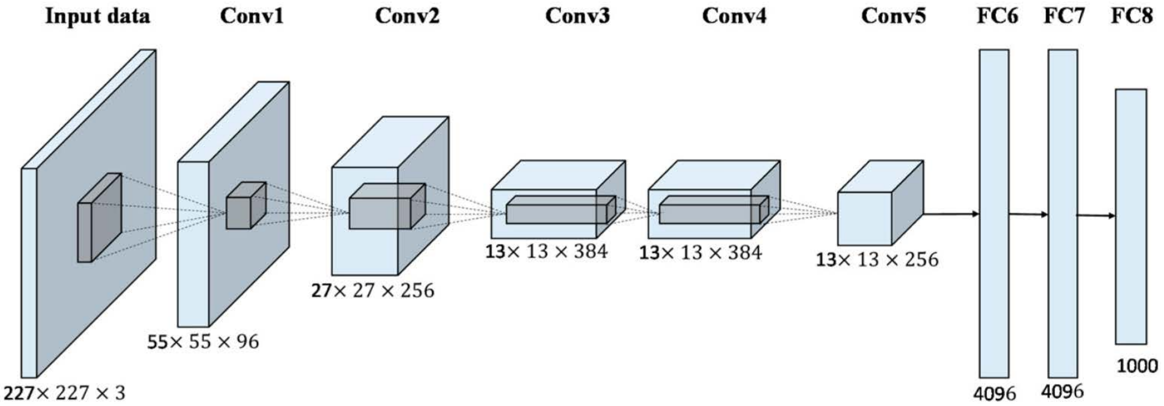


Fig. 2.3 A typical architecture of AlexNet

2.2.1.2 VGGNet

VGGNet, another highly successful CNN, was developed by the Visual Geometry Group (VGG) at the University of Oxford in 2014 [72]. It uses a combination of convolutional layers, pooling layers, and fully connected layers to classify images. VGGNet is known for its simplicity and its use of small filters (3x3) with a stride of 1 and a padding of 1. It uses ReLU as activation function and dropout as regularization technique. An example architecture of the network is presented in the Figure 2.4.

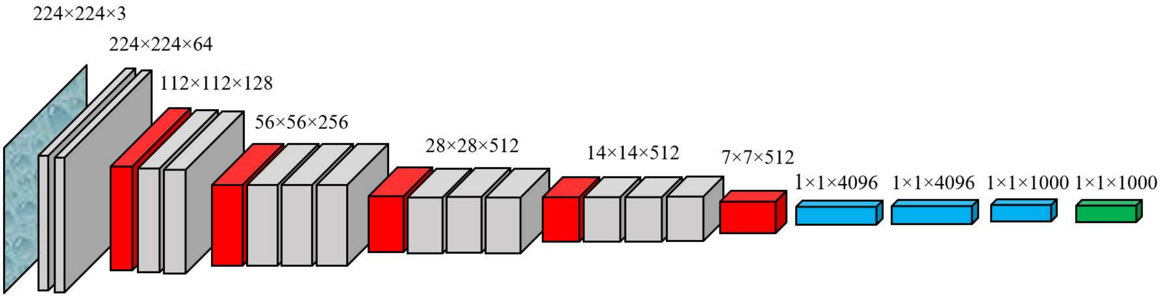


Fig. 2.4 A typical architecture of VGG network

2.2.1.3 GoogLeNet

GoogLeNet, known for its Inception module, which is a combination of multiple convolutional layers with different filter sizes, pooling layers, and fully connected layers, all in one layer, was developed by Google in 2014 [73]. This allows the network to learn multiple scales of features from the same input. GoogLeNet uses ReLU as activation function and dropout as regularization technique. An example architecture of the network is presented in the Figure 2.5.

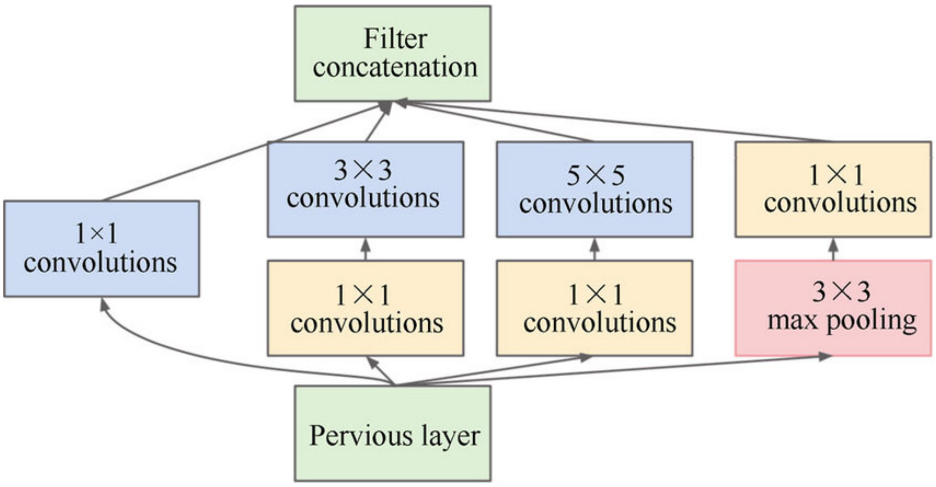


Fig. 2.5 A typical architecture of Inception module of GooggLeNet

2.2.1.4 ResNet

ResNet, developed by Microsoft in 2015, is famous for its residual block which allows the network to learn a residual function (F(x)) with reference to the layer input (x) instead of learning the original mapping (F(x)) directly [74]. This allows the network to be very deep (up to 152 layers) without suffering from the vanishing gradients problem. ResNet uses ReLU as activation function and dropout as regularization technique. An example architecture of the network is presented in the Figure 2.6.

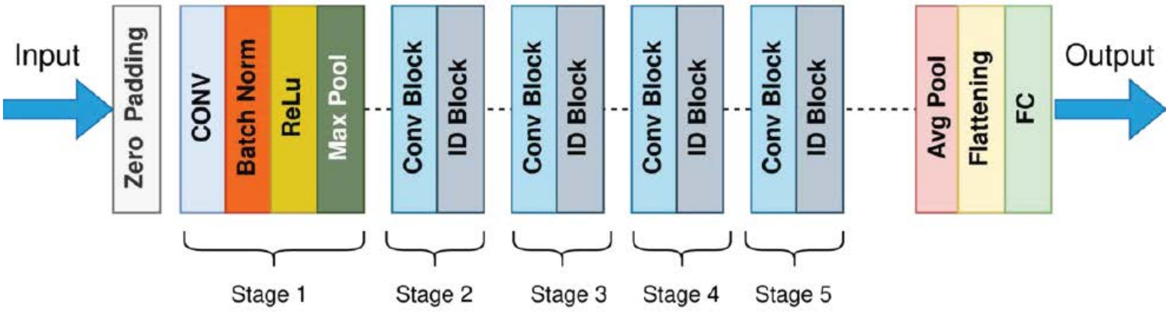


Fig. 2.6 A typical architecture of ResNet network

2.3 Deep Learning Models for Segmentation

2.3.1 Fully Convolutional Networks

Fully Convolutional Networks (FCNs) are specifically designed for image segmentation tasks [75]. They are composed of multiple layers, including convolutional layers, pooling layers, and upsampling layers.

2.3.2 U-Net

U-Net is a FCN, designed for image segmentation tasks, consists of a contracting path, that is used to extract features from the image, and an expanding path, that allows to generate the segmentation mask. The contracting path is similar to the encoder part of an autoencoder and consists of a series of convolutional and max-pooling layers. The expanding path is similar to the decoder part of an autoencoder and consists of a series of convolutional and upsampling layers [76].

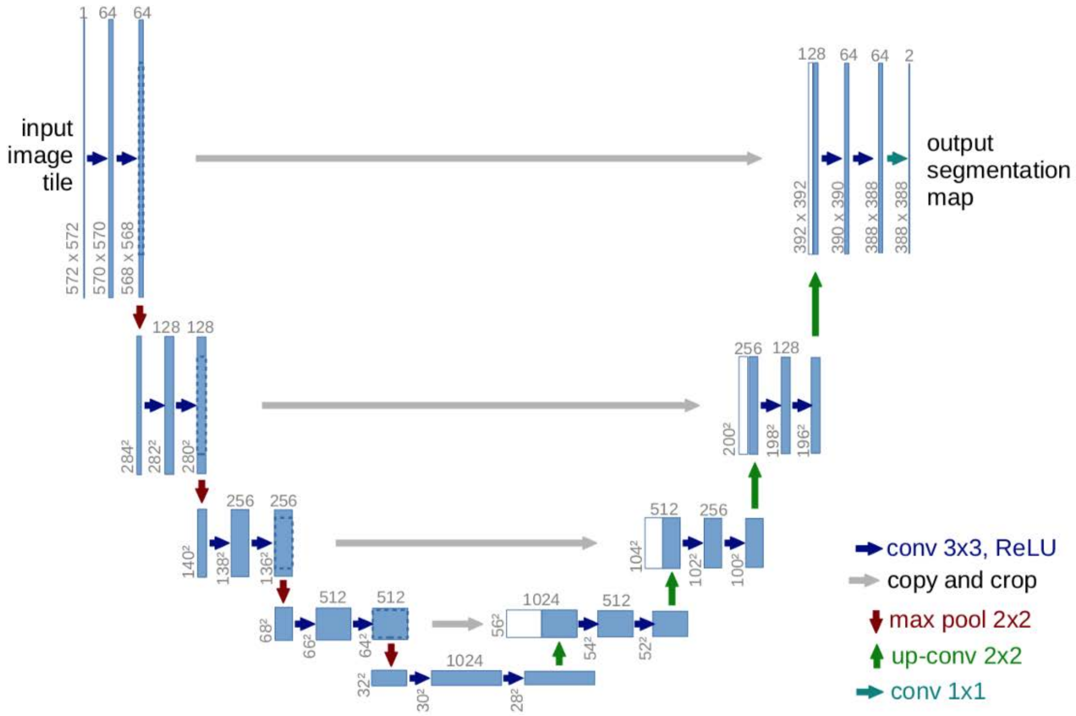


Fig. 2.7 A typical architecture of U-Net network

Additionally, U-Net uses concatenation of feature maps between the contracting and expanding path, which allows the network to propagate more context information to the deeper layers of the network. U-Net is particularly useful for segmenting images with large variations in shape and intensity. An example architecture of the network is presented in the Figure 2.7.

2.3.3 DeepLab

DeepLab is a CNN that is based on the atrous convolution which allows the network to have a larger field of view and extract more context information. This architecture uses a combination of convolutional layers, atrous convolution and fully connected Conditional Random Field (CRF) to generate the segmentation mask. Atrous convolution allows the network to have a larger field of view and extract more context information. The CRF helps in refining the segmentation results by taking into account the neighborhood information [77]. An example architecture of the network is presented in the Figure 2.8.

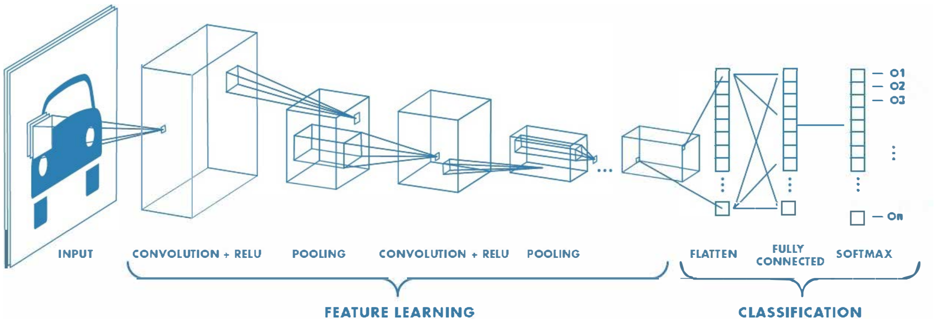


Fig. 2.8 A typical architecture of DeepLab network

2.3.4 Mask R-CNN

Mask R-CNN is an extension of the Faster R-CNN object detection network that is used for instance segmentation [78]. It uses a CNN to extract features from the image and a Region Proposal Network (RPN) to generate object proposals. It also uses a separate branch to generate a segmentation mask for each object proposal. This network combines the object detection and semantic segmentation in a single pipeline, which makes it more efficient for some tasks. An example architecture of the network is presented in the Figure 2.9.

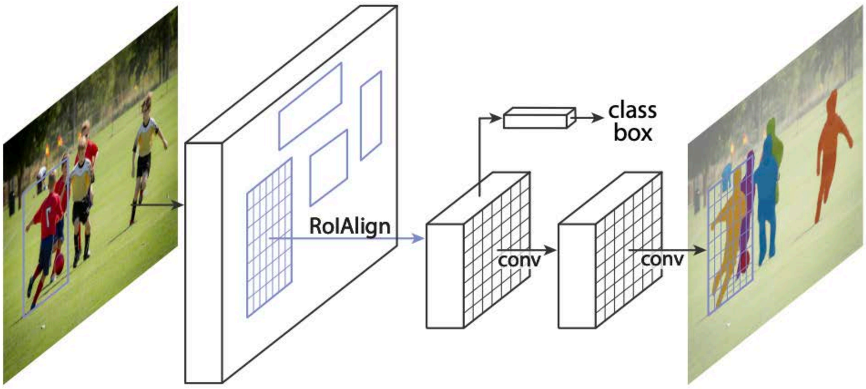


Fig. 2.9 A typical architecture of Mask R-CNN network

2.4 Deep Learning Models for Detection

2.4.1 Faster R-CNN

Faster R-CNN is a two-stage object detection method that uses a RPN to generate object proposals and a CNN to classify and locate objects within the proposals. The RPN is trained to generate object proposals that are likely to contain objects, while the CNN is trained to classify and locate objects within the proposals [79].

Region Proposal Network: The RPN is a fully convolutional network that is trained to generate object proposals. It takes an entire image as input and produces a set of object proposals, each represented by a bounding box. The RPN is trained to generate object proposals that are likely to contain objects, based on the features extracted from the image. An example architecture of the Faster R-CNN network is presented in the Figure 2.10.

2.4.2 RetinaNet

RetinaNet is a one-stage object detection method that uses a CNN to classify and locate objects within an image. It uses a combination of convolutional layers and Feature Pyramid Networks (FPNs) to extract features from the image and a separate branch to classify and locate objects within the image. RetinaNet is particularly useful for detecting small or faint objects that are difficult to detect with traditional object detection methods [80]. An example architecture of the RetinaNet network is presented in the Figure 2.11.

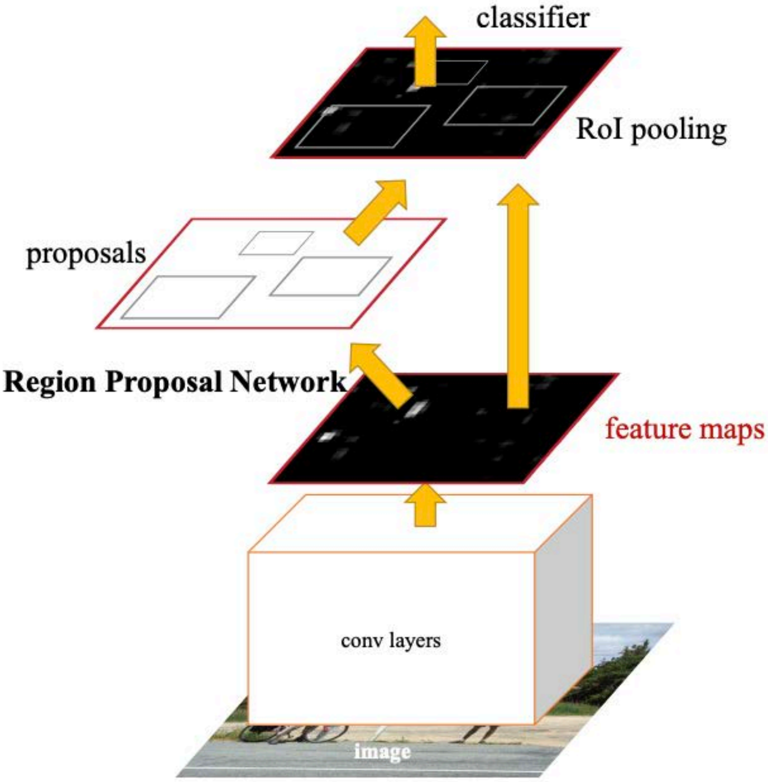


Fig. 2.10 A typical architecture of Faster R-CNN network

2.4.3 Feature Pyramid Network

FPNs are used to extract features at multiple scales. The FPNs are built on top of the convolutional layers of the network. They are used to extract features from different levels of the convolutional layers and combine them to create a rich feature map. This allows the network to detect objects of different scales [81]. An example architecture of the FPN network is presented in the Figure 2.12.

2.4.4 YOLO

You Only Look Once (YOLO) is a one-stage object detection method that uses a CNN to classify and locate objects within an image. It uses a combination of convolutional layers to extract features from the image and a separate branch to classify and locate objects within the image [82]. An example architecture of the YOLO network is presented in the Figure 2.13.

YOLO uses a grid-based prediction mechanism, where the image is divided into a grid of cells, and each cell predicts a set of bounding boxes, class probabilities, and confidence

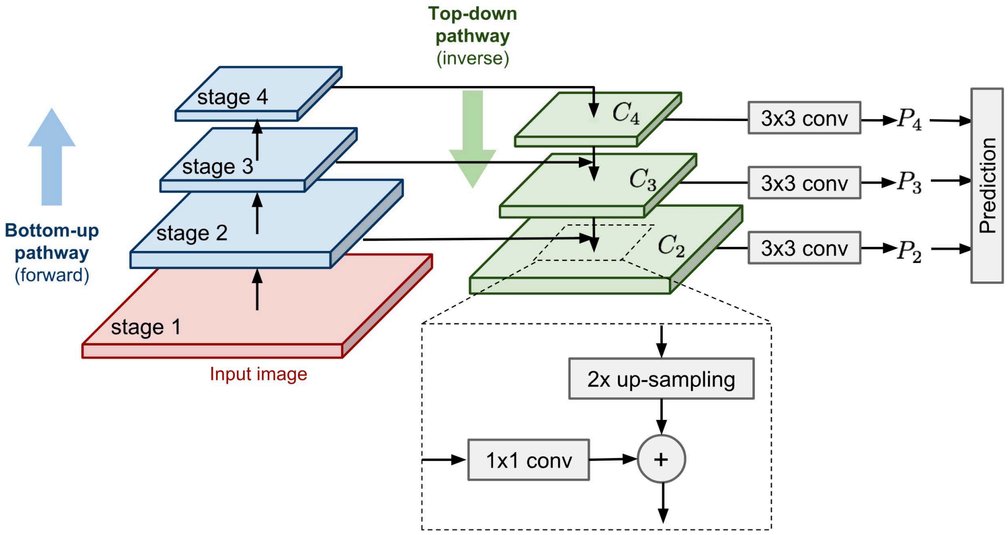


Fig. 2.11 A typical architecture of RetinaNet network

scores. The grid-based prediction allows the network to handle multiple scales and aspect ratios of objects in the same image.

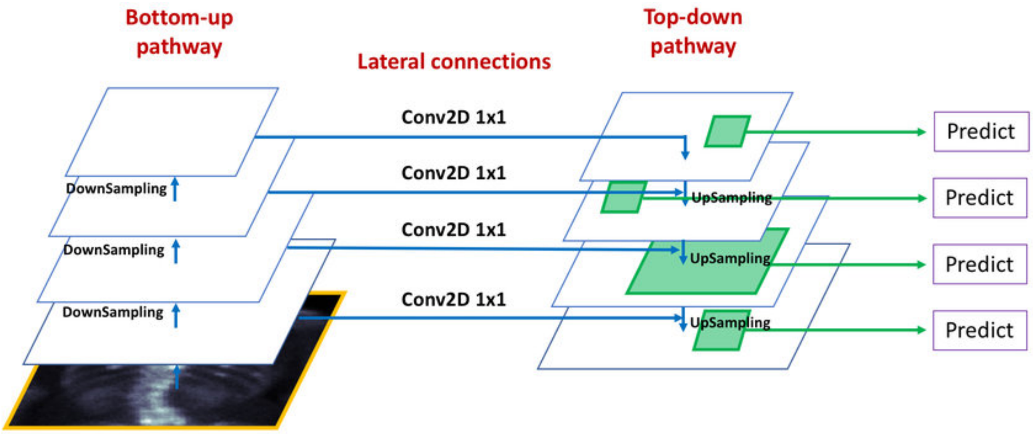


Fig. 2.12 A typical architecture of Feature Pyramid Network

2.5 Performance Measuring Metrics

In the context of medical imaging, there are several performance measuring metrics that are commonly employed to evaluate the performance of DL models. The most commonly used metrics for the classification task include accuracy, precision, recall, and f1-score based upon confusion matrix. The segmentation and the object detection tasks may have additional measuring parameters depending upon the nature and the definition of the problem. The section below highlights some commonly applied performance measuring metrics of DL in the medical imaging. Numerous performance measuring metrics are mutually adopted by classification, detection, and segmentation tasks. For the sake of concision, the repeating performance measuring metrics are explained only once (at the first appearance in the text).

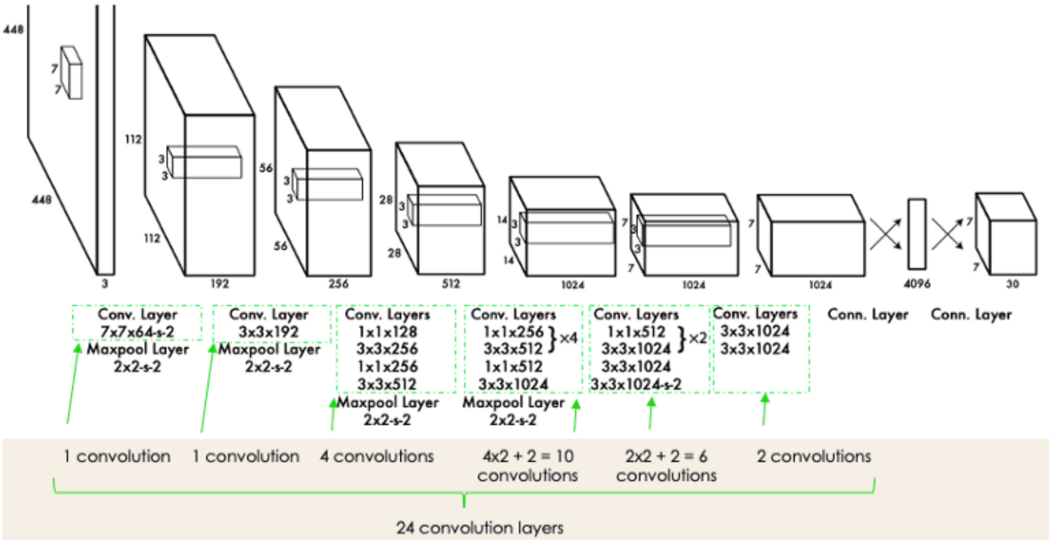


Fig. 2.13 A typical architecture of YOLO network

2.5.1 Performance Measuring Metrics for Classification

The performance of a DL classifier is generally evaluated using a confusion matrix, which provides a comprehensive representation of the model's ability to accurately predict classification outcome. The confusion matrix is a commonly used in machine learning and pattern recognition for assessing the quality of a classifier. A confusion matrix for binary classification problem is provided in the Table 2.2.

Table 2.2 *Binary classification confusion matrix.* N and P stand for Negative and Positive, respectively. TP, TN, FP and FN indicate the number of True Positives, True Negatives, False Positives, and False Negatives, respectively.

		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

A True Positive (TP) is an outcome where DL model accurately predicts the positive class. Whereas, a True Negative (TN) is an outcome of a DL model where the model correctly predicts the negative class. On the other hand, a False Positive (FP) is an outcome where the model incorrectly predicts the positive class, whereas, a False Negative (FN) is an outcome where the model incorrectly predicts the negative class.

Considering the TP, TN, FP, and FN provided in the confusion matrix, the accuracy of a binary classifier can be defined as in the Equation (2.1). Similarly, the recall, otherwise known as sensitivity, can be defined as given in the Equation (2.3), and the F_1-score, otherwise kown as Dice coefficient, is defined in the Equation (2.2). Lastly, the specificity is defined in the Equation (2.5).

2.5.1.1 Accuracy

In the context of image classification, accuracy is a measure of how well a DL model is able to correctly classify images into their corresponding classes. It is defined as the proportion of correctly classified images to the total number of images in the test set.

Mathematically, it can be represented as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

The accuracy ranges from 0 to 1, where a value of 1 indicates that the model has correctly classified all samples in the test set, and a value of 0 indicates that the model has not correctly classified any of the images in the test set.

It is important to note that accuracy alone does not provide a complete picture of a model's performance, as it does not take into account false positives or false negatives. Other metrics such as precision, recall, and F1 score are also considered when evaluating the performance of a DL model in image classification tasks.

2.5.1.2 Precision

Precision is a measure of how well a model is able to correctly classify images into their corresponding classes, among the images it predicted to be in a certain class. It is defined as the proportion of correctly classified images of a certain class to the total number of images the model predicted to be in that class.

Mathematically, it is represented as:

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

Precision is particularly useful when the cost of false positives (images predicted to be in a certain class but actually not) is high. It gives an idea of how reliable the positive predictions are. A high precision value indicates that the model has a low rate of false positives and is providing a high number of accurate positive predictions.

2.5.1.3 Recall

Recall is a measure of how well a deep learning model is able to detect all images of a certain class, among all images that are actually in that class. It is defined as the proportion of correctly classified images of a certain class to the total number of images that are actually in that class.

Mathematically, it can be written as:

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

2.5.1.4 F1 Score

F1 score is a measure that combines precision and recall to give a single metric that describes the performance of a model in image classification. It is defined as the harmonic mean of precision and recall.

Mathematically, it can be presented as:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.4}$$

2.5.1.5 Specificity

Specificity is a measure of the proportion of negative instances that are correctly identified as such by the model. It is defined as the proportion of correctly classified negative instances to the total number of negative instances. A negative instance is an image that is not part of the target class or the class of interest.

Mathematically, it can be given as:

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

Specificity is often used in conjunction with sensitivity (also known as true positive rate or recall) to evaluate the overall performance of a binary classification model.

2.5.2 Performance Measuring Metrics for Segmentation

In medical imaging, image segmentation is often employed to identify and separate different structures or regions of interest within an image. The performance of a DL model for segmentation tasks in medical imaging must be evaluated using metrics that take into account the specific characteristics of the medical images, such as the shape and size of the structures, as well as the presence of noise and artifacts.

Moreover, semantic segmentation is percieved as pixelwise or voxelwise classification problem, where additional performance measuring metric come into practice. Some of the commonly used performance measuring metrics for segmentation tasks in medical imaging are inscribed below. For all the equations listed hereunder, the letter M denotes the binarized predicted segmented volume (obtained by thresholding predicted probability maps) and A indicates the ground truth volume; the cardinality operator for a set is denoted as $|\cdot|$.

2.5.2.1 Mean Average Precision

Mean Average Precision (mAP) is a metric used to evaluate the performance of object detection models. It is the average of the average precision for each class, with the precision calculated for different Intersection over Union (IoU) thresholds. The precision is the number of true positive detections divided by the number of true positive detections plus the number of false positive detections. The IoU threshold is used to determine whether a predicted bounding box is considered a true positive or a false positive. A higher mAP value indicates better performance of the model.

The mathematical definition can be seen in the Equation (2.6) for the class H.

$$mAP = \frac{1}{H}\sum_{j=1}^{H} AP_j \tag{2.6}$$

2.5.2.2 Jaccard Index

The Jaccard Index, also known as IoU, is defined as the ratio of the intersection of the predicted segmentation and the ground truth segmentation to their union. It is important to note that JI is sensitive to the size of the sets and is not symmetric, meaning that Jaccard(A,B) is not equal to Jaccard(B,A).

Mathematically, it can written as in the Equation (2.7).

$$J(M,A) = \frac{|M \cap A|}{|M \cup A|} \tag{2.7}$$

2.5.2.3 Dice Coefficient

The Dice Similarity Coefficient (DSC) can be used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. Dice coefficient is 2 times the area of overlap divided by the total number of pixels in both the images. DSC ranges from 0 to 1, with a value of 1 indicating a perfect match between the predicted and ground truth segmentation.

Mathematically, the DSC can be represented as in the Equation (5.2).

$$DSC(M,A) = \frac{2 \cdot |M \cap A|}{|M| + |A|} \tag{2.8}$$

2.5.2.4 Volumetric Overlap Error

Volumetric Overlap Error (VOE) is a metric used to evaluate the performance of 3D object detection models. It is a measure of the difference between the predicted 3D object bounding box and the ground truth bounding box.

VOE is calculated by taking the ratio of the volume of the intersection of the predicted bounding box and the ground truth bounding box to the volume of the union of the predicted bounding box and the ground truth bounding box as reported in the Equation (4.19).

$$VOE(M,A) = 1 - J(M,A) \tag{2.9}$$

A lower VOE values indicate better performance of the model.

2.5.2.5 Others

The Tversky Index, $T_{\alpha,\beta}(M,A)$, is a generalization of the concept of overlap between$DSC(M,A)$ and $J(M,A)$, which can be explained as in the Equation (2.10).

$$T_{\alpha,\beta}(M,A) = \frac{|M \cap A|}{|M \cap A| + \alpha|M - A| + \beta|A - M|} \tag{2.10}$$

where it is worth mentioning that the $T_{0.5,0.5}(M,A)$ corresponds to $DSC(M,A)$, and $T_{1,1}(M,A)$ is equivalent to $J(M,A)$.

A more approximate indication about the relative difference between the volumes is the Relative Volume Difference (RVD), which is defined as:

$$RVD(M,A) = \frac{|M| - |A|}{|A|} \tag{2.11}$$

In image-guided surgical situations, shape and size of organ have crucial impact, therefore, accurately predicting and evaluating the shape of organ is essential. Maximum Symmetric Surface Distance (MSSD), Average Symmetric Surface Distance (ASSD), and the Root Mean Square Symmetric Surface Distance (RMSD) are key metrics in evaluating models built for surgical procedures.

To calculate these distances, a metric space (X,d) must be established. In this space, X is a 3D Euclidean space, and d is the Euclidean distance. The external surfaces of the M and A volumes, represented as $L(M)$ and $L(A)$ in X, can then be used to define a distance function, known as the one-sided Hausdorff distance, $h(L(M),L(A))$, as shown in the Equation (5.8).

$$h(L(M),L(A)) = \sup_{l_M \in L(M)} \left\{ \inf_{l_A \in L(A)} d(l_M,l_A) \right\} \tag{2.12}$$

Furthermore, the *MSSD*, also known as bidirectional Hausdorff distance, can be defined as in the Equation (5.7), whereas, the *ASSD* can be defined as in the Equation (5.5) and *RMSD* as in the Equation (2.15).

$$MSSD(M,A) = \max\{h(L(M),l(A)),h(L(A),L(M))\} \tag{2.13}$$

$$ASSD(M,A) = \frac{1}{|L(M)+L(A)|}\left(\sum_{l_M \in L(M)} d(l_M,L(A)) + \sum_{l_A \in L(A)} d(l_A,L(M))\right) \tag{2.14}$$

$$RMSD(M,A) = \sqrt{\frac{1}{|L(M)+L(A)|}} \cdot \sqrt{\sum_{l_M \in L(M)} d(l_M,L(A))^2 + \sum_{l_A \in L(A)} d(l_A,L(M))^2} \tag{2.15}$$

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

2.5.3 Performance Measuring Metrics for Detection

2.5.3.1 Intersection over Union

The IoU is another metric that is commonly used for evaluating the performance of image detection models. It is defined as the ratio of the intersection of the predicted and ground truth segmentation masks to the union of the predicted and ground truth segmentation masks. IoU also ranges from 0 to 1, with a value of 1 indicating a perfect match between the predicted and ground truth segmentation.

The mathematical notation of the IoU is given in the Equation (2.16).

$$IoU(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.16}$$

2.5.3.2 Intersection over Minimum

Intersection over Minimum (IoM) is a performance measure that compares the predicted bounding box of an object to the ground truth bounding box. The IoM is calculated by taking the ratio of the area of the intersection of the predicted bounding box and the ground truth bounding box, to the minimum of the two bounding boxes' areas.

In more detail, the IoM is calculated as reported in the Equation (2.17).

$$IoM(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \tag{2.17}$$

A value of 1 would mean that the predicted bounding box perfectly matches the ground truth bounding box, whereas a value of 0 would mean that there is no overlap between the two bounding boxes. IoM is a mostly employed when the objects in the images are of different shapes and sizes. Other commonly used metrics for object detection include IoU and mAP.

2.5.3.3 Average Precision

Average Precision (AP) is calculated by computing the precision and recall at different threshold settings for the object detector. Precision is the number of true positive detections (correctly identified objects) divided by the number of true positive detections plus the number of false positive detections (objects that were incorrectly identified as the target object). Recall is the number of true positive detections divided by the number of true positive detections plus the number of false negative detections (objects that were not identified by the detector).

The precision-recall curve is then plotted, with precision on the y-axis and recall on the x-axis. The AP is then calculated as the area under this curve.

The mathematical formulation of AP is presented in the Equation (2.18).

$$AP = \int_0^1 p(r)dr \tag{2.18}$$

The relationship between precision and recall in relation to the confidence level e can be represented using the notation $q = Q(e)$ and $t = T(e)$, where q is the precision and t is the recall at a given confidence level e.

$$Q(e) = \frac{T(e) \cdot S_i}{T(e) \cdot S_i + D(e)} \tag{2.19}$$

The value of S_i in the Equation (5.3) presents the number of objects in class i, whereas, the value of $D(e)$ denotes incorrect detections bearing the confidence e.

Part I

Deep Learning for Image-guided Surgical Applications

Chapter 3

Deep Learning Applications in Image Guided Surgery

3.1 Introduction and Background

The introduction of AI in the healthcare realm has drawn a tremendous amount of attention in recent years [83–85]. The subsequent rise of the DL has assisted the surgeons in the operating room in several different ways [1, 16]. This successful incorporation has paved the way for RAS [17]. Unlike traditional surgery, a RAS system includes a camera arm and a few other mechanical arms with surgical instruments attached. The surgeon controls the arms while seated at a computer console near the operating table. The console gives the surgeon a high-definition, magnified, 3D view of the surgical site. The purpose of RAS, as the name suggests, is not to replace the surgeons and physicians but to assist them, in order to achieve higher proficiency in security and safety of the undergoing patients in preoperative, intraoperative, and postoperative surgical procedures [18, 19].

Image driven DL methods for robotic surgery have already taken care of the instrument detection and segmentation [20, 21], gesture recognition [22], workflow analysis [23], skill assessment [24], and many more [25–28] to facilitate the semi-autonomous RAS.

Moreover, the development of a fully autonomous image-guided surgical system, where the direct involvement of the surgeon is seldom required, is foreseeable task for the DL models. The surgical procedures go through several complicated scenes and contain artefacts and performance variances [24]. Additionally, the blur images and videos generated by camera are often misinterpreted and mislabelled by physicians and AI systems, because of the presence of smoke, shade of tools, plasma stains and vessels [36–39].

Before the advent of the modern image modality capturing systems, the surgeons mostly relied on simple cameras and naked eyes to study the internal behaviour of the organs. Today, the most relied imaging modalities include X-rays, CT, MRI, US, and PET [11, 12].

However, even the modern imaging modalities required intensive preprocessing and feature engineering [13]. Thanks to the DL, this laborious, time consuming, and cost intensive task is no more as tedious as heretofore. Moreover, the basic underlying principle of the DL mimics the (functionality of) biological neuron, connects with a complex layered structure, learns from generalization, and keeps the neuron-associated weights updated. One of the most powerful models of the DL is believed to be the CNN. The introduction of the CNN can be traced back to early 1960s [14], which has led to the development of several highly efficient diagnostic systems [15].

The DL has ultimately proven the enormous success in MIS systems. The very first RAS system i.e. da Vinci surgical system, introduced in the year 2000, has successfully performed around $1,594,000$ surgical procedures in 2021 [29] with an increase of 28% from the previous year ($1,243,000$ in 2020) and is expected to perform $12-15\%$ more in the following part of the year.

The MIS reduces the post-surgery trauma, minimises the hospital stay, improves recovery, and avoids potential risk of contagion [30]. The extreme difficulty of indirect surgical operation leads to the development of instrument tracking, gaze estimation, gesture and trajectory recognition, hand-eye coordination, organ and smoke detection, and depth and pose estimation systems [86–92].

Furthermore, the research in the DL based image driven RAS systems is expanding and also the availability of recent datasets, i.e. Johns Hopkins University and Intuitive Surgical Inc. Gesture and Skill Assessment Working Set (JIGSAWS), Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cholec80, and ATLAS Dione [93–96] has boosted the interdisciplinary synergies of biomedical engineers and physicians.

Several recent survey articles span the medical domain [97–99], however lack the DL part in the technical aspects. All the reviewed technical surveys consider a specific application of deep learning and image processing in the robotic-assisted surgery, such as: surgical phase recognition [100], skill assessment [101], registration [102], tool tracking or segmentation and detection phases [103, 104]. For instance, the study by Rivas et al. [105], published in 2021, considered merely one article published post 2020, and mostly emphasized on available surgical datasets and future of robotics. Another article by Unberantha et al. [106] surveyed 2D/3D image registration in workflow analysis. The study was limited to the CNNs and has not incorporated robotic part and surveyed only one particular subdomain in the

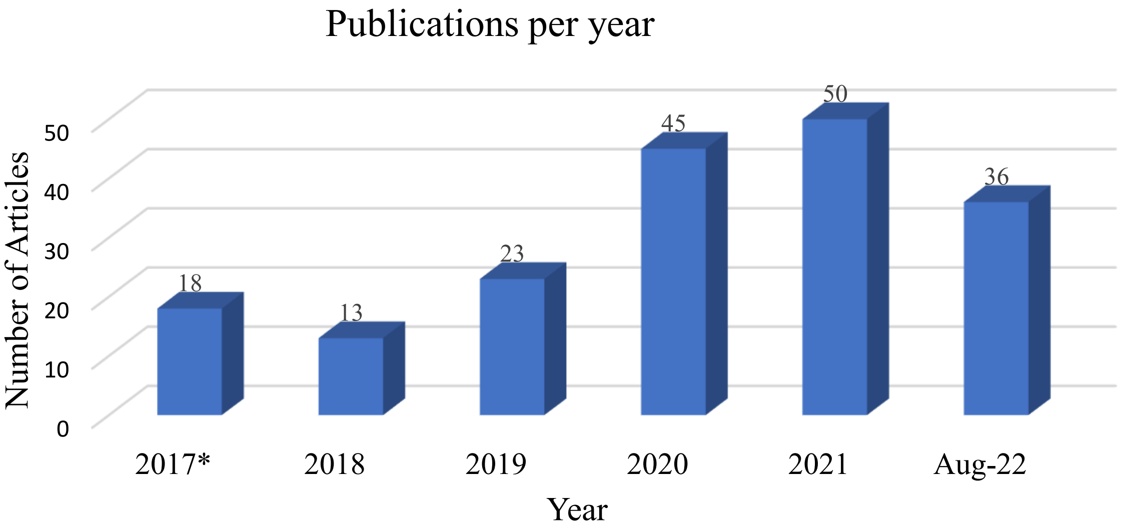


Fig. 3.1 Number of publications per year in the filed of image-guided RAS. The * represents the articles including 2017 and the previous years.

RAS. In author's opinion, an updated survey that deals with and encircles all the possible applications and aspects is required by the research and medical communities, especially for the new researchers in the field to have the possibility to see the big picture. Another aspect that encouraged the author to perform a new survey study concerns with the exponentially growing number of publications in the field, as depicted in the Figure 3.1, since the existing survey articles are pretty old. Authors in [64] have included merely three articles published post 2019, and in last couple of years, a great number of worthy articles have contributed to the domain. Additionally, the main focus of the survey remained limited to tool tracking. From recent studies, it can clearly be observed that image and video guided DL based robotic surgery survey dates back quite a few years, and in the meantime, a huge number of studies have been published on the topic. Therefore, a comprehensive updated survey is missing that can accommodate the DL part and the clinical part, considering image and video driven robotic surgery in light of the recent advancements.

After the comprehensive analysis and thorough survey, the selected papers are classified into 4 different classes, i.e. Surgical Tools, Surgical Processes, Surgical Surveillance, and Skills/Performance Assessment. Each of these classes are further subdivided, and the details can be found in the below presented sections. The full text analysis revealed that majority of the articles included in the survey are published in year 2020 and 2021 as shown in the Figure 3.1. The most frequently used DL method and dataset are CNN and JIGSAWS, whereas, the tool segmentation and detection are most studied subcategories within RAS.

3.2 The Literature Search and Survey Methodology

The following section describes the literature survey methodology adopted in this study. Initially the literature search and inclusion and exclusion protocols are provided, followed by the article selection process. Moreover, the objectives and the results are also illustrated for the conducted review. Finally, the survey classification layout is presented.

3.2.1 Literature Search

A thorough literature search is performed on Scopus® database to select the relevant articles for review and analysis. The conducted search is confined to the literature published in English language. To retrieve the optimised results, the combination of the keywords is used interchangeably with slight modifications over repetitive iterations of web search. The specific query used for the final search is: ("deep learning*" OR "deep-learning*" OR convolution* OR "deep networks*" OR "neural network*") AND (surg*) AND (robot*). The survey study is conducted on the published articles (including those accepted and available online) until August 31, 2022.

3.2.2 Inclusion and Exclusion

The inclusion criteria span the image driven DL models used in any type of robotic surgery. The search query is constrained to computer science, engineering, biomedical engineering, and medical disciplines. Only the published articles are included without considering the books, seminars, doctoral symposiums, and talks. Any article that goes beyond the aforementioned limits, any study not tackling surgery or a part of surgery, the articles related to only engineering side of the robot, and the articles related to only medical side of the surgery (i.e. no intervention of DL) come under the exclusion criteria. The Figure 3.2 illustrates the stages of the inclusion and exclusion process flow with number of studies included and excluded at each phase.

3.2.3 Article Selection

Initially, the titles of the articles and the venues of the publications (i.e. publishing authority and domain) are used to decide the relevancy on a general scale. In the further stages, the abstracts are reviewed, and the contents of each study are skimmed to limit the number of articles to the decided realm for the survey study. Finally, the full-text review is performed,

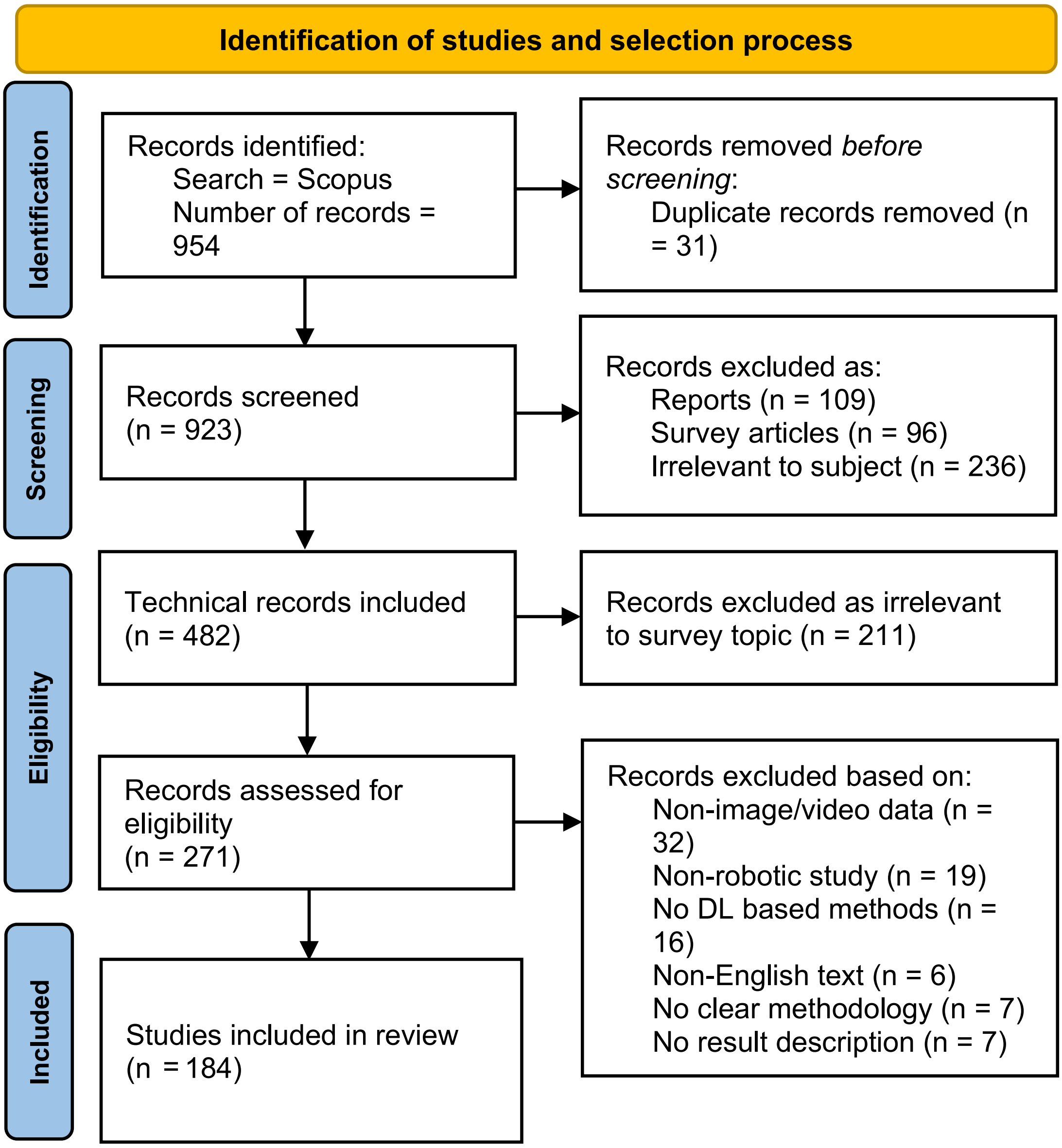

Identification of studies and selection process
Identification
Screening
Eligibility
Included
Records identified:
Search = Scopus
Number of records = 954
Records removed *before screening*:
Duplicate records removed (n = 31)
Records screened (n = 923)
Records excluded as:
Reports (n = 109)
Survey articles (n = 96)
Irrelevant to subject (n = 236)
Technical records included (n = 482)
Records excluded as irrelevant to survey topic (n = 211)
Records assessed for eligibility (n = 271)
Records excluded based on:
Non-image/video data (n = 32)
Non-robotic study (n = 19)
No DL based methods (n = 16)
Non-English text (n = 6)
No clear methodology (n = 7)
No result description (n = 7)
Studies included in review (n = 184)

Fig. 3.2 The flow diagram of the paper selection and pruning process according to the recommendations of the PRISMA method.

and the appropriate articles are selected for further proceedings. The Figure 3.2 illustrates the selection stages of the survey which is performed under the recommendations of Preferred Reporting Items for Systematic Review and Meta Analysis (PRISMA).

3.2.4 Targeted Objectives

The primary objective of the case study is to systematically analyse and summarize the recent contributions in the field of image-guided robotic surgery accounting the advancements of the DL. Generally, the study is conducted on RAS systems and specifically on image based RAS systems.

Additionally, the study aims to comprehensively state DL methods, the future of surgical robotics, and the challenges to achieve the autonomous surgery. Finally, the secondary objectives include the introduction of currently available surgical datasets, the legal and ethical issues, and the limitations of the existing systems.

3.2.5 Results

The aforementioned query resulted in a total number of 879 articles and the minor changes (upto date search) in the query showed 75 additional results. After the first check i.e. title, relevancy, and venue, a sum of 482 articles are found appropriate. Another 211 articles are discarded amid irrelevancy to the scope of the survey. At each of the stages, a considerable number of the articles is rejected and at the final stage, 184 articles are tagged eligible to the purview of the study, therefore, 184 articles out of total 954 are appended in this study as shown in the Figures 3.1 and 3.2.

3.2.6 Classification of the Case Study

After the thorough analysis, the relevant studies are found to be greatly overlapping that can be organised in numerous different topologies. However, the careful inspection resulted into four groups, each of which is further classified into several subgroups as depicted in the Figure 3.3. This classification includes: a) Surgical Tools, b) Surgical Processes, c) Surgical Surveillance, and d) Surgical Performance/Assessment.

The Surgical Tool section is further subdivided into Tool Detection and Tool Segmentation sections, the Surgical Processes includes Gesture Segmentation, Trajectory Segmentation, and Tissue Segmentation categories. The Surgical Surveillance is segregated in Surgical Planning, Phase & State Estimation, and Activity Recognition phases, and the last but not least, Surgical Skill Assessment and Surgical Workflow Recognition come under the Surgical Performance/Assessment group.

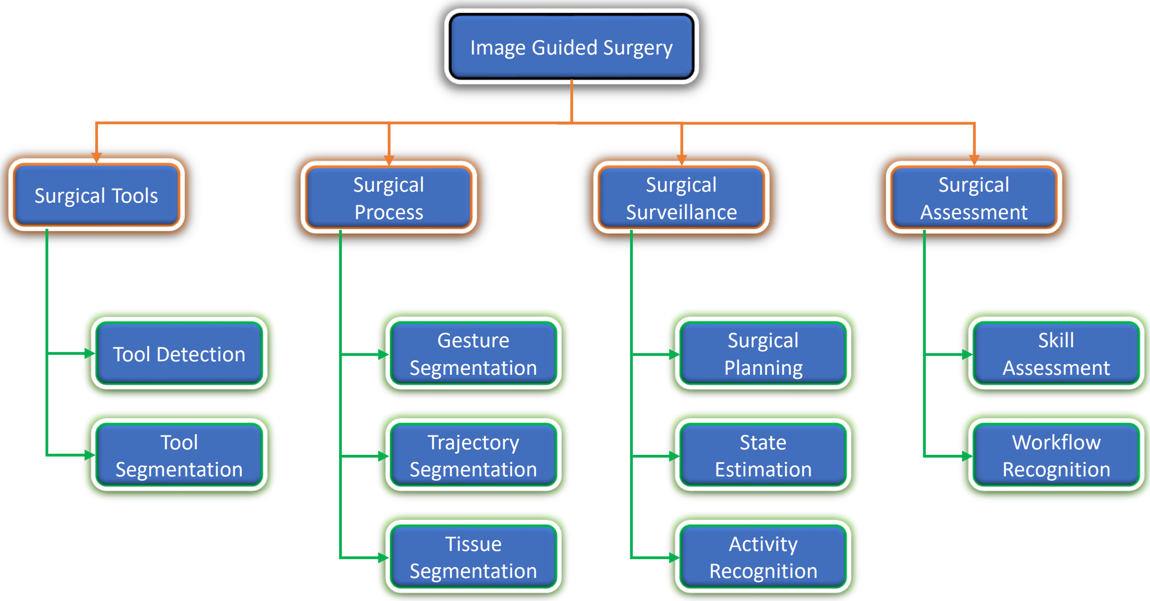


Fig. 3.3 The taxonomy of the case study.

3.3 DL Assisted Image Guided Surgery

In the previous decade, DL methods brought tremendous amount of success, especially in image-guided CAD systems as illustrated in the Figure 3.1. This enormous triumph gave birth to the idea of image driven DL models in the field of robotic surgery. The availability of large amount of image data, the less complicated operational facilities, and the DL algorithms' performance on image dataset are major knocks towards autonomous surgery.

Based on the results of case study, the articles are classified into four categories, each of which are subdivided into further groups. These categories include Surgical Tools, Surgical Processes, Surgical Surveillance, and Surgical Performance. As the names suggest, the divisions are fundamental and encircle the most relevant parts of surgical scenarios in computer-assisted autonomous and semiautonomous surgical systems.

An additional fifth category named *Others* is provided for the applications that either do not fall in any of the aforementioned categories or the number of found articles were fewer. The section below inscribes all categories in detail.

3.3.1 Surgical Tools

Surgical tools are the most important actuators in surgery because they are responsible for performing interventions; however, keeping track of surgical instruments requires real-time knowledge of the pose and the movement of the tool. Literature suggests numerous tool localisation techniques embracing electromagnetic tracking [107], kinematic [108], optical tracing [109], and image-guided detection [110] among others [111]. Unlike other approaches, image driven surgical instrument localisation offers attractive benefits including the knowledge of pose and motion, and does not require instrument design modification [104].

The section below contains the literature studied in this study about image-based surgical tool detection and segmentation which are most studied areas in robotic surgery with an average of around 40% of the total publications encompassed in this study.

3.3.1.1 Tool Detection

This section includes the articles that deal with the presence of surgical instruments in surgical videos. Among the articles studying surgical tools, 55% are about detection and/or recognition, whereas the other half of the articles belongs to the forthcoming subsection of tool segmentation. The CNN is the most applied DL method followed by Long Short Term Memory (LSTM), RNN and autoencoder architectures (Figure 3.4). The CNN model and the variants, with few modifications in the underlying architecture in some cases, yielded better performance in [86, 111–114, 117–120, 122, 123, 125–130, 132–136, 138, 141, 143, 144], whereas autoencoders, RNN, LSTM, and GAN formed another notable synergy [96, 115, 116, 121, 124, 131, 137, 139].

The reason behind CNN being the most applied architecture lies in the ability of multiple tool detection and localisation which traditional ML models have not been sufficiently successful at [111]. Among the tool detection, the articles employing public datasets revealed an accuracy range of 89-100%, whereas, the precision and Dice values vary greatly. The in-house datasets are incorporated by 18 studies achieving an overall accuracy range of well above 90% except one study ([86]) that managed to reach around 85%.

The Endoscopic Vision (EndoVis) challenge [94] and m2cai16-tool datasets [96] are most widely used followed by the ATLAS Dione for the task of tool detection. Furthermore, accuracy is top used performance measuring metric with precision and Area Under the Curve (AUC) being the other most important evaluation parameters. The more details about the year of publication, objective/s, data description and performance outcome can be found in the Table 3.1.

Table 3.1 The summarised results of the articles dealing with the tool detection task

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Microsurgery Tool Tracking	LeNet	RMIT	1171 Frames 480 × 640 Pixels 3 Surgeries	Accuracy: 99.13%	2016	[112]
Line Tracking	CNN	In-house	1000 Frames 2500 Training Images	Accuracy: 99.7%	2017	[113]
Surgical Tool Detection	CNN YOLO	M2CAI16-Tools Dataset	10 Procedures 2532 Frames	Recall: 80.62% Precision: 84% mAP: 72.26	2017	[114]
Tool Landmark Detection	Encoder-Decoder CNN	In-house	10 Sequences 1500 HD Images	RMSE: 25.479 μm	2017	[115]
Robotic Tool Detection	Faster RCN RPN	ATLAS Dione	10 Surgeons 99 Videos total 854 × 480 22,467 Images	Precision: 91%	2017	[96]
Tool Joint Detection	3D FCNN U-Net	EndoVis UCL dVRK	10 Videos 1083 Frames 720 × 576 Resolution 8 Videos 3075 Frames	DSC: 88.6% DSC: 86.9% Dice: 85.1%	2019	[116]
Tool Localization & Detection	ResNet-18 50-152 AlexNet VGG-16	cataRACT	50 Videos 10 Min & 56 Sec Duration	AUC: 0.65 0.68 0.64 0.58	2019	[117]
Guidewire Tip Tracking	U-Net	In-house	11 Videos 11268 Frames	Dice: 88.07% IoU: 85.07%	2019	[118]
Needle Localization	ResNet-18 RetinaNet	In-house	19,200 Images 512 × 1024 Resolution	Accuracy: 99.2%	2019	[119]
Surgical Tool Detection	Hourglass VGG-16	ATLAS Dione EndoVis	99 Video 10 Surgeons 22467 Frames 1083 Frames 720 × 576 Resolution	mAP: 91.60% mAP: 100%	2019	[120]
Surgical Tool detection	CNN VGG-M	M2CAI16-Tools Dataset	10 Procedures 2532 Frames	Accuracy: 89%	2019	[121]
Instrument Detection	YOLO9000 CNN	M2CAI16-Tools Dataset	10 Procedures 2532 Frames	mAP: 84.7	2019	[122]
Surgical Tool Detection	ResNet-18 ResNet-101 Hourglass-104	ATLAS Dione EndoVis	99 Video 10 Surgeons 22467, 1083 Frames 720 × 576 Resolution	mAP: 98.5% mAP: 100%	2020	[123]
Needle Detection	LSTM CNN	In-house	NA	100% TPR	2020	[124]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Instrument Detection	VGG-16 CNN	ATLAS Dione	10 Surgeons 99 Videos total 854 × 480 22,467 Images	Precision: 90.08%	2020	[125]
Surgical Tool Navigation	ResNet-18	In-house	4500 Image	Accuracy: 137 µm	2020	[126]
Instrument Detection	ConvNet	EndoVis	300 Frames 640 × 480 Size 400 Frames 640 × 480 Size	Accuracy: 91.2% Accuracy: 75%	2020	[127]
Needle Insertion Tracking	U-Net	In-house	30 Porcine Eyeballs 300 Training Images 1024 × 640 Pixels	Errors: 7.4 µm 10.5 µm 3.6 µm	2020	[128]
Object Recognition	CNN	In-house	5670 Images 3968 × 2976 Size	Accuracy: 98%	2020	[129]
Needle Detection	Faster RCN	In-house	27 Videos 9 Subsets	Precision: 89.2% IoU: 73.9%	2020	[130]
Tool Presence Analysis	Multitask RCN LSTM	Cholec80	80 Videos 13 Surgeons 854 × 480 Resolution	mAP: 89.1% F1 Score: 87.4%	2020	[131]
Tool Tracking	GAN	EndoVis15	3 Videos 44s Long	Accuracy: 95.2%	2020	[132]
Surgical Tool Detection	CNN	Cholec80 EndoVis	80 Videos 13 Surgeons 854 × 480 Resolution 1083 Frames 720 × 576 Resolution	mAP: 91.6% mAP: 100%	2021	[133]
Tool Tip Detection	RetinaNet YOLOv2	In-house	2310 Frames 9 Videos 640 × 480 Resolution	Recall: 1.000 Precision: 0.733 F1 Score: 0.846 Recall: 0.864 Precision: 0.808 F1 Score: 0.835	2021	[134]
Needle Detection & Segmentation	CNN NN	In-house	2D US Images Terason uSmart 3200 T NexGen US system 22-gauge 0.7 mm diameter 80 mm length	Accuracy: 99.7% Precision: 86.2% Recall: 89.1% F1-score: 0.87	2021	[135]
Instrument Tracking	TernausNet-11 TernausNet-16 MobileNet-V3 ShuffleNet-V2	In-house	1846 Images 640 × 480 Size	Accuracy: 85.87%	2021	[86]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Object Detection	YOLOv4	M2CAI16-Tools Dataset	10 Procedures 2532 Frames	Recall: 79.1% Precision: 96.7%	2021	[136]
Tool Detection & Segmentation	YOLACT	In-house	5,319 Frames 70 Videos 1920 × 1080 Resolution	Accuracy: 91.2% Precision: 56.5% Dice: 48.2%	2021	[137]
Needle Tracking	NN YOLO-v3	In-house	778 & 332 Images	Accuracy: 1.98mm	2021	[138]
Instrument Detection	Faster R-CNN	In-house	5085 Images	IoU: 0.825 Recall: 0.950 Precision: 0.950	2022	[139]
Instrument Tracking	YOLO-v4	In-house	6243 Images	Accuracy: 95.12%	2022	[140]
Tool Detection	YOLO-v5	In-house	20 Videos 7500 Frames	Precision: 89.5–91.4%	2022	[141]
Instrument Triplet	SIR-Net	EndoVis18	16 Videos 1280 × 1024 Pixels 8 Instruments	Average Precision: 0.6515	2022	[142]
Micro-robot Detection	VGG Net	In-house	15000 Ultrasound Images	Accuracy: 0.95 & 0.93	2022	[143]
Object Detection	ResNet-101 Back Projection	Data Generation from Video	380 Images	Accuracy: 94.12% Recall 86.23%	2022	[144]
Object Detection	YOLO-v4 Faster-RCNN MobileNet EfficientDet	In-house	196.55 Minute Videos 870 Images	mAP: 29.3, 22.2, 23.4, 33.6 F1 Score: 75.86, 82.34, 82.49, 93.50	2022	[111]

3.3.1.2 Tool Segmentation

Surgical instrument segmentation is different from surgical instrument detection in terms of binary, semantic, and instance segmentation. Generally, tool detection either looks for the presence of any tool (recognition) or the location of a particular tool (tool tip or landmark detection), whereas the segmentation distinguishes (i.e. segments) the tools from other organs and also differentiates among numerous tools. It involves the individual identification of each instrument within an image. As mentioned in the above subsection, tool segmentation is second most common researched field inside the image-guided RAS. Instead of only tool presence recognition, numerous articles focus on the type of tool available in the surgical procedure with semantic segmentation [103, 145–170]. A noteworthy point arises when articles dealing with organ/object segmentation (see Section 3.3.5) also consider tool

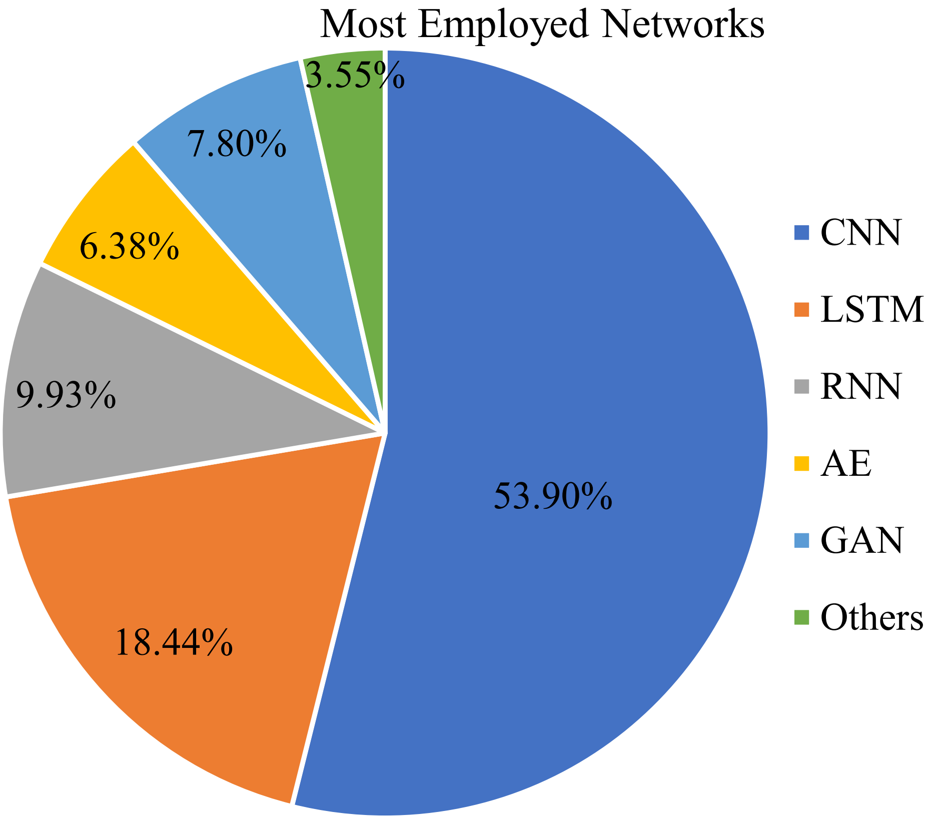


Fig. 3.4 Percentage of the articles with respect to employed DL model. Articles which adopted two or more models are counted accordingly.

segmentation [171], therefore forming another interconnected relation between two different but relevant tasks.

Additionally, real-time instrument segmentation has gained ample amount of attention in recent studies [172–177]. Numerous authors also consider the semantic segmentation of a part of a particular instrument, such as tool tip segmentation, guide-wire segmentation and needle segmentation [178–180].

Likewise, semantic segmentation by using unsupervised DL methods is another growing concept [181, 182]. The data provided by EndoVis robotic instrument segmentation challenge [96] is most frequently used dataset for segmentation, whereas the Dice score is common performance measuring metric. Out of total 38 studies, merely 4 studies incorporated in-house datasets and 2 assimilated both in-house and public datasets. The description of the input, results and other relevant information is provided in the Table 3.2.

Table 3.2 The summarised results of the articles dealing with the instrument segmentation task

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Surgical Tool Segmentation	CNN RNN Auto Encoder-Decoder	EndoVis16	4 Videos 45-seconds Each 720 × 576 Resolution 25 Frames	Accuracy: 93.3% Jaccard Index: 82.7%	2017	[147]
Automatic Instrument Segmentation	U-Net TernausNet-11 LinkNet	EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	IoU: 83.60 Dice: 90.01 Time: 184	2018	[103]
Guidewire Tip Segmentation	Faster R-CNN	In-house COCO PASCAL VOC	22 Sequences 2D X-ray Images 1080 × 1080 Pixels	Precision: 0.532 F1 Score: 0.939	2018	[178]
Binary Segmentation	ResNet-18 FNN	EndoVis17	8 Sequences 1280 × 1024 Resolution 225 Frames 8 Sequences 75 Frames	IoU: 0.764	2019	[145]
Instrument Segmentation	CNN	EndoVis17	225 Frames 8 Surgeries	Dice: 0.916 Hausdorff: 11.11 Specificity: 0.989 Sensitivity: 0.928	2019	[172]
Semantic & Instance Segmentation	U-Net	EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	Dice: 90.20%	2019	[146]
Realtime Instrument Segmentation	MobileNet-v2	EndoVis17 Cata7	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frame Videos	Dice: 96.91% IoU: 94.10% Dice: 58.30%	2020	[175]
Object Extraction for Instrument Tracking	U-Net ResNet-18	M2CAI16-Tools Dataset	10 Procedures 2532 Frames 1280 × 720 Pixels	Accuracy: 100%	2020	[152]
Surgical Instrument & Workflow Recognition	Bayesian AlexNet LSTM	Cholec80	80 Videos 13 Surgeons 25 fps 854 × 480 Resolution	Bipolar: wMAE: 0.76 pMAE: 0.96 Scissors: wMAE: 0.51 pMAE: 0.76	2020	[153]
Tool Segmentation	FCNN ResNet-50 U-Net	In-house	14 Videos 300 Frames 720 × 576 Pixels 4200 Annotations	IoU: 81.80/7.74	2020	[176]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Synthetic Image Segmentation	U-Net	In-house	160 Frames 5 Videos	mIoU: 0.235 mIoU: 0.458 mIoU: 0.729	2020	[154]
Multi-Angle Instrument Segmentation	TernausNet-16 VGG-16	Sinus-Surgery-C Dataset	10 Videos 5-23 Minute 320 × 240 Resolution	mDSC: 90.2±0.% mIoU: 85.6±1.2%	2020	[155]
Unsupervised Learning for Instrument Segmentation	CycleGAN DRN	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	mIoU: 0.732	2020	[181]
Ultrasound Needle Segmentation	LinkNet	In-house	996 Images 102 Videos 3 fps	IoU: 41.01% Dice: 56.65% F1 Score: 36.61% RMS: 13.3	2020	[179]
Instrument Segmentation	GAN	EndoVis18 EndoVis17	8 Sequences 225 Frames 19 Sequences	Accuracy: 76.29%	2020	[156]
Tools Collision Avoidance using Segmentation	U-Net	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	MAE: 0.126 ± 0.08 mm	2020	[160]
Instrument Segmentation	ResNet-18	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	IoU: 0.852 Time: 11.8 ms IoU: 0.729 Time: 11.8 ms.	2020	[161]
Image-to-Image Translation for Instrument Segmentation	GAN CNN	Sinus-Surgery-C	10 Videos 320 × 240 Resolution	mDSC: 82.7 mIoU: 75.5	2020	[159]
Unsupervised Instrument Segmentation	Vanilla U-Net	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	IoU: 0.71 Dice: 0.81	2020	[182]
Instrument Segmentation	CycleGAN	Cholec80 EndoVis15	80 Videos 13 Surgeons 854 × 480 Resolution 300 Images & 6 Videos	Dice: .80	2020	[162]
Surgical Instrument Segmentation	CycleGAN U-Net	EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	Dice: 92.8% IoU: 84.7%	2021	[148]
Real-Time Instrument Segmentation	LSTM	EndoVis18 EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos 19 Sequences 15 Training Videos 4 Test	mDice: 61.03% mIoU: 53.89% mDice: 77.53% mIoU: 67.50%	2021	[173]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Multi-Instance Segmentation	Encoder-Decoder CNN	EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	mAP: 0.481 ± 0.099 mIoU: 0.657	2021	[149]
Real-Time Instrument Segmentation	VGG MobileNet ResNet	UW-Sinus-Surgery-C/L ROBUST-MIS	10 Videos 5-23 Minute 320 × 240 Resolution 3 Videos 12-66 Minute 1920 × 1080 Res	mDSC: 3.1% 9.5% mIoU: 3.3% 10.7%	2021	[174]
Instrument Segmentation	U-Net	EndoVis17 ISBI2018	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	mAUC: 0.6819 IoU: 83.70% Dice: 90.24%	2021	[150]
Surgical Tool Segmentation	GAN U-Net	EndoVis17	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	IoU: 0.867 Dice: 0.924	2021	[151]
Robot Positioning Using Instrument Segmentation	YOLOv3 ResNet	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	IoU: 86.6%	2021	[157]
Real-Time Instrument Semantic Segmentation	MobileNet-v3	EndoVis17	8 Sequences 225 Frames 2 × 300 Frames Videos	IoU: 69.74% Dice: 79.88% Hausdorff: 11.36	2021	[177]
Guide Wire Segmentation	MobileNet U-Net	In-house	1050 Images 1440 × 1560 Pixels	Accuracy: 97.81%	2021	[180]
Instrument Segmentation	Modified CNN	MICCAI 2018	15 Videos	IoU: 0.4354 Accuracy: 0.9638	2022	[163]
Instrument Segmentation	Modified CNN	EndoVis & In-house	10 & 20 Videos 4 & 5 Scenarios 720 × 576 & 1920 × 1080 Pixels & 5000 Frames	Average Accuracy: 93.31%	2022	[163]
Instrument Segmentation	U-Net & VGG-16	Hamlyn's & Proprietary	1920 × 1080 Pixel 8 × 255 Frames	IoU: 0.708 & 0.826	2022	[164]
Instrument Segmentation	SurgiNet & MobileNet-v2	EndoVis 2017 & CataIS	10 & 9 Videos 3000 & 2671 Images 7 & 11 Instruments	Mean IoU: 89.14% & 63.30%	2022	[165]
Surgical Tool Segmentation	DenseNet ResNet-18	Kvasir-Instrument EndoVis17	590 Videos 2 × 300 Frames Videos	Mean IoU: 0.900	2022	[166]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Instrument Segmentation	Modified CNN	EndoVis & In-house	10 & 20 Videos 4 & 5 Scenarios 720 × 576 & 1920 × 1080 Pixels & 5000 Frames	Average Accuracy: 93.31%	2022	[163]
Instrument Segmentation	U-Net & VGG-16	Hamlyn's & Proprietary	1920 × 1080 Pixel 8 × 255 Frames	IoU: 0.708 & 0.826	2022	[164]
Instrument Segmentation	SurgiNet & MobileNet-v2	EndoVis 2017 & CataIS	10 & 9 Videos 3000 & 2671 Images 7 & 11 Instruments	Mean IoU: 89.14% & 63.30%	2022	[165]
Surgical Tool Segmentation	DenseNet ResNet-18	Kvasir-Instrument EndoVis17	590 Videos 8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames	Mean IoU: 0.900	2022	[166]
Semantic Segmentation	Modified U-Net	D^{sia} D^{por}	48702 Images	Structural Similarity: 77.53 IoU: 74.92 Dice: 85.63	2022	[167]
Semantic Segmentation	RNN Attention Module Encoder-Decoder	kvasir-instrument UW-Sinus-Surgery-C/L	590 Images 768 × 576 Resolution 4345 Images 5-23 Minute Videos 320 × 240 Resolution	Dice: 96.27 mIoU: 92.82	2022	[168]
Tool Tip Segmentation	U-Net	UW-Sinus-Surgery-C/L	8360 Images 5-23 Minute Videos 30 fps 256 × 256 Resolution	mDice: 0.9522 mIoU: 0.9088	2022	[169]
Instrument Segmentation	Mask R-CNN CNN Swin-Transformer	EndoVis 2017	8 Sequences 225 Frames Test Set 8 × 75 Frames 2 × 300 Frames Videos	mIoU: 0.5873 mIoU: 0.7408	2022	[170]

3.3.2 Surgical Processes

Surgical process is a nontechnical parent terminology induced to explain those sub-tasks of MIS which are not directly relevant to incision but lead to the understanding and developing the next generation autonomous medical robotic systems. The surgical process section is subdivided into three sections, i.e. Gesture Recognition, Trajectory Segmentation, and Tissue Segmentation, based on the contribution of the authors towards the field.

In addition, numerous studies employ these terminologies interchangeably, however, considering the in-depth analysis, the suturing task cannot be confused with unique movement of surgical instrument. The former can be perceived as series of analogous gestures, whilst latter can be described as the movement in and around a particular region.

3.3.2.1 Gesture Segmentation

Surgical gesture recognition has been perceived in several different overlapping contexts i.e. path planning, needle positioning, continuous tip detection, etc. and therefore, numerous DL methods have been applied in respective perspectives. The gesture segmentation is generally implemented for suturing tasks, therefore, it is a cumbersome job because of the similarity and repetition of analogous suturing steps.

The gesture recognition can be either live or in-vitro environments, however for suturing tasks, it is broadly available as in-vitro experiment in the literature [22, 27, 183–189]. The live suturing task involves risks and requires close consideration and high costs, therefore fewer studies adopt live suturing [190–192].

The LSTM model is adopted by five out of total thirteen studies, whereas, CNN and RCN are other most applied models by several authors. Moreover, all thirteen articles used accuracy to measure the performance of the DL models along with other metrics and nine out of thirteen studies incorporated JIGSAWS [93] dataset and three employed in-house datasets (among these two studies used both JIGSAWS and other datasets also).

However, the outcome of these research studies is evident that fusion of different types of data (i.e. video and kinematic data) yields better accuracy as compared to only image/video data. The further technical details extracted from the gesture segmentation works are enlisted in the Table 3.3.

3.3.2.2 Trajectory Segmentation

The task of trajectory segmentation also involves the motion analysis and pattern recognition of the involved surgical tools. Similar to the gesture segmentation task, the in-vitro experiments are generally used [194–203]. To improve the results of segmentation, authors also incorporate the kinematic data along with the video and image data [194, 199–202, 204]. The kinematic data has particular importance because it leads to the learning from demonstration.

Likewise, not only the thread detection but also knot tying and path planning are largely associated with trajectory segmentation tasks [195, 196, 198, 201, 202, 205, 206]. It is worth mentioning that instead of using one single architecture, authors used combination of DL architectures to produce better performance results (e.g. CNN and LSTM). The Table 3.4

Table 3.3 The summarized results of the surgical gesture recognition and segmentation articles.

Procedure	**DL Model**	**Dataset**	**Input Datatype**	**Input Data Description**	**Results**	**Ref**	**Year**
Laparoscopy	NN	In-house	Video Data	7 Gestures 2 Tools	Accuracy: 100% & 80%	2006	[187]
In-vitro Suturing	RCN LSTM	JIGSAWS	Video & Kinematic	11 Gestures 14 Sequences	Accuracy: 71% & 67%	2019	[185]
In-vitro Suturing	3D CNN	JIGSAWS	Video & Kinematic	39 Videos 11 Gestures	Accuracy: 84.3%	2019	[186]
In-vitro Suturing	CNN	JIGSAWS	Image Data	10 Gestures 39 Sequences	Accuracy: 81.67%	2020	[183]
Live Suturing	RCN LSTM	JIGSAWS	Video Data	10 Gestures 39 Sequences	Accuracy: 85.5%	2020	[190]
Live Suturing	CNN	JIGSAWS	Video Data	10 Gestures 39 Sequences	Accuracy: 90.1%	2020	[191]
In-vitro Suturing	3D CNN LSTM	JIGSAWS	Image Data	10 Gestures 39 Sequences	Accuracy: 76.3%	2020	[22]
Prostatectomy Suturing	AlexNet LSTM ConvLSTM	In-house	Video & Kinematic	2395 & 511 Videos 5 Gestures	Accuracy: 78% & 62%	2021	[27]
Suturing Tasks	LSTM	JIGSAWS In-house	Video & Kinematic	12 Gestures	Accuracy: 75%	2021	[192]
In-vitro Suturing	SD-Net	JIGSAWS	Video Data	10 Gestures 39 Sequences	Accuracy: 90.5%	2021	[184]
Prostatectomy Suturing	TCN	JIGSAWS RARP-45	Video & Kinematic	39 & 45 Videos 12 & 7 Gestures	Accuracy: 86.8% & 81%	2022	[193]
Hand Gesture	Modified MobileNet-v2	In-house Jester	Video	30 Subjects 2*30 Gesture 148,092 Videos 7 Gestures	mAP: 96.82%	2022	[188]
In-vitro Suturing	ResNet-50 TCN	JIGSAWS	Video & Kinematic	39 Videos 10 Gestures	Accuracy: 89.8%	2022	[189]

highlights the salient features of the trajectory segmentation task and the employed DL models.

3.3.2.3 Tissue Segmentation

The tissue segmentation appears to be the third largest studied task in this study comprising around 12% of the total articles (see Table 3.5). This section comprises all the studies

Table 3.4 The summary of the results of the trajectory segmentation publications

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Trajectory Segmentation	Deep CNN VGG AlexNet	JIGSAWS	10 Gestures 39 Sequences	Silhouette Score: 0.733±0.056, 0.716±0.097	2016	[199]
Pattern Cutting	Deep Reinforcement Learning	In-house	NA	IoU: 0.833	2017	[198]
Trajectory Planning	CNN	In-house	NA	Accuracy: 91.25%	2017	[201]
Trajectory Segmentation	Convolutional Auto-Encoder	JIGSAWS	10 Gestures 39 Sequences	Accuracy: 78.2% Accuracy: 92.1%	2018	[200]
Trajectory Segmentation	Dense Convolutional Encoder-Decoder Network	JIGSAWS	28 Videos 8 Video 11 Voices 38 Video 10 & 19 Voices	Accuracy: 70.8% Accuracy: 62.1%	2018	[194]
Trajectory Generation	CNN	In-house	60 & 10 Cable Images	IoU: 0.754 IoU: 0.583	2018	[204]
Thread Tip Detection	CNN LSTM with RNN	NA	1278 & 1215 Labeled Images	Precision: 99.63 Recall: 98.89	2019	[195]
Pedicle Screw Path Planning	CNN based 3D U-Net	NA	21 Spinal CT Images	Dice: 95.55 Jaccard: 91.92 MSE: 1.340	2019	[196]
Trajectory Planning	NN	LumSeg SpiSeg xVertSeg	105 CT Scans	Positioning Error ± Std: 2.37±0.97	2020	[205]
Trajectory Generation	DNN, FC DenseNet	LumSeg SpiSeg xVertSeg	105 CT Scans	Positioning Error ± Std: 2.37±0.97	2021	[197]
Path Planning	3D U-Net	In-house MICCAI	15 & 8 CT Scans	Accuracy: 93%	2022	[206]
Path Planning	GAN CNN LSTM	In-house	NA	Accuracy: 72.94%	2022	[202]
Motion Prediction	TCN Attention Module	In-house	33 + 25 Subjects	RMSE: 1.02	2022	[203]

performed on tissues including vessel segmentation, edge detection, healthy and cancerous tissue classification, uncertainty inference segmentation, and tissue retraction [207–218].

The applicability of tissue segmentation spans the liver to the brain to the kidney to the lungs and to several other organs [218–225]. It also includes the binary classification of healthy and cancerous tissues [213, 215, 226].

The aforementioned division of the segmentation phase is placed under the same category because of the overlapping interest and the main task involved in the article. The individual categories can be assumed in a study that only encircle the tissue segmentation task regardless of the input type. The basic reason behind including the smaller categories (even with fewer published papers found) is the significant contribution discussed in the robotic surgery field.

The U-Net architecture is the most applied network among all studies with being adopted by eleven out of seventeen articles, as evident in the Table 3.5. The U-Net is often applied in conjunction to the other networks including LSTM, GAN [224] and other variants of CNN [208, 210, 213, 214, 216, 221, 223]. Because of the unavailability of large scale public datasets, a significant majority of the studies (13 out of 18) incorporated in-house dataset. Considering the proprietary datasets, authors have used varying performance measuring metrics including accuracy, IoU, Dice and AUC. The further insights about the datasets, performance, and the DL techniques are provided in the Table 3.5.

3.3.3 Surgical Surveillance

The increasing introduction of biomedical images facilitates the surgical surveillance and the navigation during the surgical process. This section surveys the articles that monitor the surgical situation with respect to the patient and the ongoing procedure.

3.3.3.1 Surgical Planning

Surgical planning is largely considered as preoperative planning, where the steps are performed in advance in order to pre-visualise the intervention. The application has a large benefit in emergency situations and war field areas where reaching the hospital requires time. The vessel detection and needle insertion are prominent and attractive applications, and surgical robots are made to achieve timely fashioned aid [227]. The recent trends in surgical planning have shown great interest in the 2D, 3D, and 4D model construction for the interventional guidance [227–235]. Due to the unavailability of the large amount of data for preoperative planning, majority of the articles rely upon in-house data [228–231, 233–235] which includes CT scans and MRI scans [228–230, 232, 233].

Table 3.5 The concise results of the tissue segmentation papers

Procedure	DL Model	Dataset	Data Description	Results	Year	Ref
In-vitro Experiment	FFNN	In-house	144 Samples	IoU: 0.911	2016	[222]
Laproscopy	CNN	In-house	2050 Video Frames	Dice≥0.95	2017	[219]
Ex-vivo Experiment	AlexNet VGG19 Inception-v3	In-house	250000 Frames	Accuracy%: 99.47 99.52, 99.71	2018	[213]
Spondylectomy	GAN, DCNN U-Net, VNet	DeepLesion	282 Images 3863 Images	IoU: 0.9584	2019	[224]
Laproscopy	U-Net, LinkNet SegNet, FCN	EndoVis19	Videos 1 & 2 EndoVis19 12 Videos	IoU: 78.31	2020	[214]
Arthroscopy	U-Net	In-house	18278 2D Images	Dice: 0.87%	2020	[215]
Arthroplasty	U-Net U-Net++	In-house	3868 Images	Dice: 0.64%	2020	[208]
Lobectomy	DNN U-Net	In-house	1080 Images 62 Minutes Video	Accuracy: 83.4% ± 3.3%	2020	[210]
Arthroscopy	Bayesian CNN	In-house	16973 17944 Images	AUC: 90.0 AUC: 89.2	2020	[211]
In-vivo Experiments	U-Net	EndoVis17	150 Images from da Vinci	IoU: 0.3	2020	[220]
Arthoplasty	GoogLeNet	In-house	500 Images	Accuracy: 97.8%	2020	[225]
In-vivo & Ex-vivo Experiments	SVM RF CNN	In-house	53 Patients 67893 In-vivo 89695 Ex-vivo	ROC-AUC: 0.88	2020	[226]
Nephrectomy	FCNN 2D U-Net 3D U-Net NephCNN	Nephrec9	8 RAPN videos 1871 Frames	Dice: 71.76%	2021	[221]
Prostatectomy	U-Net ResNet MobileNet	In-house	5 Videos 15570 Images	IoU: 0.894	2021	[223]
Gastrectomy	U-Net	In-house	33 Videos 30 fps	Mean Recall: 0.606 Mean Dice: 0.549	2021	[216]
Abdominal Surgery	U-Net LSTM	FlapNet	2736 Sequences	Accuracy: 83.77%±2.18%	2021	[207]
Neurosurgery	Modified U-Net	In-house Proprietary	25 fps 40, 34, 41 Frames 224x288 Pixels	Dice: 0.97, 0.86 0.87 0.77	2022	[217]
In-vivo Experiments	CNN	In-house	9059 Images 17777 Annotations	Accuracy: 0.95	2022	[218]

As can be seen in the Table 3.6, the majority of the authors use Dice performance measure. Along with the surgical planning, the articles [228, 234, 235] also deal with the depth estimation, motion detection and path planning, which are also crucial parts of RAS. The articles are enlisted under the surgical planning section considering their major contribution towards the category.

3.3.3.2 Phase and State Estimation

The surgical phase and state estimation subgroup a particular surgical process into several chunks of phases and establish a system that recognises the phase or state of the surgical process on a given set of inputs. The task is largely applied in several domains including cholecystectomy, endovascular, and esophagectomy procedures [131, 236–243], however, it appears to have less generalization for other types of surgical procedures.

Dissimilar to the surgical planning, the availability of the data for phase recognition makes majority of the articles rely upon public datasets [131, 236, 238–240, 243] which include kinematic data with images and videos. The CNN and LSTM are two most frequently applied methods with accuracy being the top performance measuring parameter. Phase and state estimation are directly related to identifying the status of the process at certain time, therefore, the accuracy is used by almost all studies to evaluate the performance of employed DL models. The Table 3.7 provides further details on the phase and state estimation studies.

3.3.3.3 Activity Recognition

The automatic surgical activity recognition before the surgical procedure and in the operating room during the surgical intervention gained considerable attention in the recent past [245, 246]. The surgical activity recognition involves the real-time followup of the procedure under consideration. The integration of CNN and LSTM networks helped surgeons draw reasonable conclusions, however, the unavailability of large datasets has led to use the pretrained DL models [247–250]. The pretrained DL models are highly trained on ImageNet dataset [251].

Recently, surgical activity recognition has been considered the essential part of surgical planning and state estimation, therefore, as can be observed from the above two subsections, the activity recognition comes under the umbrella of surgical procedure planning. Apart from the video and image data, the kinematic data is integrated to generate better results, whereas, the accuracy is considered the reasonable performance measure. The Table 3.8 highlights the main concepts, methods, datasets, and other relevant information about the articles addressing the surgical activity recognition tasks.

Table 3.6 The summary of the studies related to the surgical planning

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Preoperative Model Tract	ANN	In-house	10 Patients 3 Months	Kapa Index: 0.78	2016	[230]
Syndrome Planning	CNN LSTM	In-house	MRI & 3D US 12 Monochorionic Twin Pregnancies	Dice: 0.85 ± 0.06 & 0.79 ± 0.05 Jaccard: 0.73 ± 0.10 & 0.66 ± 0.08 AUC: 0.88 ± 0.06 & 0.84 ± 0.03 Sensitivity: 0.77 ± 0.10 & 0.73 ± 0.07 Specificity: 1.0 & 0.99	2018	[233]
Stent Graft Modeling & Planning	U-Net	In-house	78 Images	Distance Error: 1–3 mm	2018	[234]
Intraoperative Liver View	CNN VGG16	In-house	2016 Liver View 9504 Live Liver View 100 Patients	mAP: 85.9%	2020	[235]
4D Guidance & Construction	U-Net	In-house	600 Scans	Dice: 0.5 mm Precision: 0.794 ± 0.065 Recall: 0.803 ± 0.047 Z-coverage: 0.790 ± 0.087	2021	[229]
3D Reconstruction of Wound Edge	ANN	LumSeg xVertSeg SpiSeg	Camera Images Kinematic Info	MSE: 0.67	2021	[231]
Vascular Access Planning	YOLO-v3	In-house	19000 Images	Mean Time: 53 ± 36s	2021	[227]
Laminae Planning	SegReNet DenseSeg FC-DenseNet	In-house	10 Scans 15 Scans 10 Scans	Dice: 96.38%	2021	[232]
Automatic Ablation Planning	LeNet-5	In-house	20 OCT Volumes	Precision: 1.16 mm Error 0.74 mm	2021	[244]

3.3.4 Surgical Performance/Assessment

The performance and skill assessment of surgeon in a surgical procedure is one of the most crucial aspects of autonomous robotic surgery because it lays down the steppingstone for computer-aided autonomous systems. The skill evaluation of the surgeons along with the

Table 3.7 The summary of the studies dealing with the task of phase and state estimation using DL models

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
State Perception	CNN	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 96%	2019	[237]
Action Recognition	ResNet Encoder-decoder	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 81.71% Edit Score: 91.74 F1: 80.08	2020	[236]
Tool Presence Analysis	RCN LSTM	Cholec80	80 Videos	mAP: 89.1% F1 Score: 87.4%	2020	[131]
Pull State Recognition Needle Detection	YOLOv3 CNN	In-house	15505 Images	Accuracy: 72.4% Accuracy: 93.2%	2020	[241]
Phase Recognition	CNN RCN	Bypass40	40 Surgical Procedures	Accuracy: 90.9 ± 3.2 Precision: 85.6 ± 4.5 Recall: 84.0 ± 4.2 F1 Score: 84.2 ± 4.2	2021	[240]
State Estimation	DNN U-Net LSTM	HERNIA 20	20 Inguinal Hernia Repair Surgeries on da Vinci	Accuracy: 80.4%	2021	[238]
State Estimation	LSTM	HERNIA 20 RIOUS+ JIGSAWS	20 Inguinal Hernia Repair Surgeries on da Vinci 40 Uson da Vinci 39 Sequences 8 Users 10 Gesture Classes	Accuracy: 82.7% Accuracy: 89.5% Accuracy: 85.6%	2021	[239]
Phase Recognition	Temporal CN	In-house	31 Patients' Videos	Accuracy: 84%	2022	[242]
Phase Recognition	Temporal CN ResNet-50	M2CAI16 Cholec80	10 Procedures 2532 Frames 1280 × 720 Pixels 80 Videos	Accuracy: 91.8 ± 8.1 Precision: 90.3 ± 6.4 Recall: 90.0 ± 6.4 Jaccard: 81.2 ± 5.5	2022	[243]

workflow recognition from a surgical video allows to compute the dexterity and precision. Therefore, building a robotic surgical system requires the beforehand understanding of the skill assessment and flow during the procedure.

3.3.4.1 Skill Assessment

The manual skill assessment and skill development monitoring of the doctors, surgeons, and trainees is burdensome tasks and requires a great deal of time and expertise. This usual task

Table 3.8 The summary of the articles incorporating DL methods for the task of activity recognition

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Action Segmentation	VGG & AlexNet	JIGSAWS & 50 Salad	39 Sequences 8 Users 10 Classes 50 Instances 25 Users 2 Trials	Accuracy: 74.22% Accuracy: 72%	2016	[249]
Activity Recognition	RP-Net & Inception-v3	In-house	100 Videos 12 Tasks Each	Precision: 80.9% Recall: 76.7%	2018	[250]
Activity Recognition	3D ConvNet & LSTM	In-house	400 Surgical Videos 103 Procedures 8 Surgeons	Precision: 88%	2020	[245]
Surgery Type Recognition	CNN & LSTM	Laparo425	425 Videos 9 Surgeries	Accuracy: 75%	2020	[246]
Surgical Action Recognition	Deep CNN & pretrained CNN	Lapgyn4DS	30,682 Frames 8 Actions 500 Surgeries	Accuracy: 99.20% AUC: 99.12%	2020	[247]
Surgical Activity Recognition	Multitask CNN & ResNet-18	CholecT40	40 Videos Cholec80 128 Triplets	Accuracy: 89.7% Action Triplet Recognition: 24.78%	2020	[248]

comes under the responsibility of expert doctors, which is not only arduous but also prone to errors. The automatic surgical skill evaluation for the RAS is indispensable.

The surgical skill assessment and skill level assessment are widely realised using CNN models driven by video and kinematic data [252–258]. The JIGSAWS [93] is most common dataset for skill evaluation since it provides both video frames and kinematic data (Table 3.9). All the studies included in this section evaluated the models on accuracy and/or AUC. The skill assessment also handles the instrument tracking in some cases where the dexterity of surgeons carries extreme importance. The other performance measures, dataset description, and DL models are provided in the Table 3.9.

Table 3.9 The summary of the articles incorporating DL methods for the task of skill assessment

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Surgical Skill Assessment	CNN	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 100%	2018	[252]
Subjective Skill Assessment	3D ConvNet	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 95.1% Accuracy: 100%	2019	[253]
Surgical Skill Evaluation	Mask-RCN RPN	BABA In-house	84 Frames 454 Frames 1766 Frames	RMSE: 3.52 mm AUC: 1 mm Accuracy: 83%	2020	[254]
Skill Level Assessment	1D CNN LSTM	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Macro Avg: 0.9917 & 0.0802 Micro Avg: 0.9844 & 0.0442	2020	[255]
Surgical Performance Assessment	DNN	In-house	254 Videos 2 Simulation Exercise	Accuracy: 83.1% Accuracy: 80.8%	2021	[256]
Objective Skill Assessment	RNN	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 95.74% Accuracy: 83.54 Accuracy: 84.23% Accuracy: 81.58%	2021	[257]
Surgical Skill Assessment	Domain-Adapted Model	JIGSAWS	39 Sequences 8 Users 10 Gesture Classes	Accuracy: 96.27%	2022	[258]

3.3.4.2 Workflow Recognition

Surgical workflow analysis describes the steps involved during the surgical interventions. Automatizing the surgical workflow has great importance in the modern operating room. Autonomous workflow recognition is vital in developing computer-aided autonomous and semiautonomous surgical frameworks. These systems have the ability to supervise the surgery within operating room by scheduling the tasks and resources and providing the seamless assistance to clinicians.

The interpretation of the recorded video of the surgical procedures requires expertise, focus, and huge amount of time. The technological advances automatically extract the valuable information by analysing the videos. The cholecystectomy procedure videos are largely used in literature to understand the workflow and surgery type recognition [153, 259–262], followed by nephrectomy [246, 263–265]. The CNN and the LSTM are common methods used to study the workflow in the videos [153, 246, 259–267] because of their high performance. Eight out of ten studies used accuracy as a performance measuring metric along with others. The workflow recognition also overlaps with the future state prediction and phase recognition [265, 268]. The Table 3.10 contains the other necessary and relevant details about the section of the study.

3.3.5 Others

This section describes the articles that do not come under the major categories, however their contribution towards the image-guided RAS is not negligible. The other reason of this distinctive but amalgam section formation is the small number of found publications for the relevant subclass. Therefore, this section highlights the objective, contribution, DL methods adopted, and other significant details of each study.

In [269], the authors proposed dual neural network based models for organ recognition and presence or absence of internal organ in endoscopy image data. The second neural network model testifies the presence of the organ on series of images on the live screen. The in-house generation of small dataset resulted in 92% of the accuracy with only 200 randomly selected images for testing.

Similarly, the segmentation of organs [171, 270, 271] and tissues is also well studied task in RAS [214]. The Mask-RCN and CNN based YOLO, U-Net, TernausNet, LinkNet, and SegNet are applied on famous EndoVis Challenge and in-house datasets. The aforementioned articles coincide with the tool detection and segmentation category because of the segmentation of tool during organ detection.

In another similar study, the volume of organ segmentation in intraoperative guidance is studied using U-Net and V-Net architectures [272]. Two publicly available datasets namely VISCERAL and SLiver07 datasets are used in this study and 12.6% and 6.2% IoU for the aortic and liver segmentation are achieved.

Table 3.10 The summarised results of the flow recognition articles

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Workflow Recognition	RNN CNN ResNet LSTM	Cholec80 MICCAI 2016	80 Videos 13 Surgeons 25 fps 854 × 480 Resolution 41 Videos 25 fps 1920 × 1080 Resolution	Accuracy: 90.7% Accuracy: 92.4%	2018	[261]
Self Supervised Workflow Analysis	ResNet-50	Cholec80	80 Videos 13 Surgeons 25 fps 854 × 480 Resolution	Accuracy: 92.7 ± 4.3 Recall: 87.0 ± 4.0 Precision: 87.6 ± 5.3 F1 Score: 84.6 ± 5.4	2018	[262]
Surgery Type Recognition	CNN LSTM	Laparo425	425 Videos 9 Surgeries	Accuracy: 75%	2019	[246]
Automatic Workflow Analysis	CNN LSTM	In-house	9 Videos 24 Hz 82:49 ± 37:54 Minutes Length	Accuracy: 100% Precision: 100% Recall: 100%	2019	[264]
Flow & Context Recognition	RNN LSTM	In-house	9 Videos 24 Hz 82:49 ± 37:54 Minutes Length	Accuracy: 74.29% Accuracies: Clamping: 100% Dissection: 83% Suturing: 87% Drainage: 100% Ultrasound: 43%	2019	[263]
Surgical Workflow Recognition	ResNet-50	In-house	8 Videos 1920 × 1080 Resolution 30 fps	Accuracy: 0.9482% Loss = 0.0765	2020	[259]
Optical Flow Prediction	U-Net	Inhouse RIDE	66 Patients 700 Sequences 1920 × 1080 Resolution 25 fps	s-EPE: 2.6 (2.6) l-EPE: 14.7 (7.9) Grid EPE 15.8 (7.9)	2020	[266]
Surgical Workflow Analysis	RNN ResNet-50	Cholec80	80 Videos 13 Surgeons 25 fps 854 × 480 1920 × 1080 Resolution	Accuracy: 85.73 ± 6.96 Precision: 82.94 ± 6.20 Recall: 85.04 ± 5.15 Jaccard: 69.96 ± 8.83 F1 Score: 82.08 ± 6.45	2020	[260]

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Instrument and Workflow Recognition	Bayesianc AlexNet LSTM	Cholec80	80 Videos 13 Surgeons 25 fps 854 × 480 Resolution	Bipolar: wMAE: 0.76 pMAE: 0.96 Scissors: wMAE: 0.51 pMAE: 0.76	2020	[153]
3D Workflow Analysis	U-Net ResNet-50	In-house	9 Videos 20 Minutes Length	Mean IoU: 80%	2022	[268]
Workflow Detection	CNN Mask R-CNN	M2CAI2016	10 Procedures 2532 Frames 1280 × 720 Pixels	mAP: 96.8	2022	[265]

The gaze estimation has an important role in tele-robotic surgery, however fewer studies have incorporated image data to support autonomous surgical systems. The [273] proposed a dense CNN architecture to control the surgical robot using gaze estimation. The local dataset generated from camera images is manually labelled in nine different gaze directions. The accuracy of the direction based gaze estimation reached 90%.

Smoke detection is another less common but growing application within the robotic surgery domain. The authors in [274–276] use intraoperative images to detect and remove smoke using U-Net, GoogLeNet, and CycleGAN. The obtained results are considerable and detailed in the Table 3.11 for further reading.

An image based gauze detection and segmentation approach is proposed by Sanchez et al. [277] where pretrained models are employed to test the novel dataset of 4003 video provided by authors. The InceptionV3, MobileNetV2, and ResNet-50 managed to reach an accuracy value of 77.68%, 75.67%, 90.16%, respectively.

3.4 Publicly Available Datasets

The impressive outcomes of the DL are underscored by the large amount of available datasets because of the intrinsic nature of neural networks. Neural network based models require training which follows the principle of more the data is available, better the results are. During this case study, a decent number of articles is found to have used publicly available datasets, however, numerous studies are conducted on proprietary datasets. The publicly available datasets do not only provide benchmark for DL model development but also support in evaluation and comparison of several models.

Table 3.11 The summary of the articles incorporating DL methods for miscellaneous tasks in robotic surgery realm

Objective	DL Model	Dataset	Data Description	Results	Year	Ref
Gaze Estimation to Control Robot	Dense CNN	In-house	9 Gaze Direction	Accuracy: 90%	2018	[273]
Organ Recognition	NN	In-house	200 Images Endoscopy Records	Accuracy: 92%	2019	[274]
Volume Segmentation in CT Images	3D DCNN U-Net V-Net	VISCERAL SLiver07	20 CT Volumes 20 CT Volumes	IoU: 12.6% IoU: 6.2%	2020	[269]
Pixel-wise Smoke Detection	U-Net	Hamlyn Cholec80 m2cai16-workflow	21000 Images 91 Videos 10 Procedures 2532 Frames	MSE: 0.002 ± 0.001	2020	[214]
Smoke Detection	GoogLeNet	In-house	4500 Images 8 Surgeries 30000 Images	ROC: 0.98 AUC: 0.92-0.97	2020	[272]
Organs Segmentation	U-Net TernausNet LinkNet SegNet FCN	EndoVis 2019	12 Videos 25 Hz 960 × 540 Resolution	IoU: 78.31% Time: 28.07 ms	2021	[275]
Object Detection	YOLO-v4	In-house	5 Videos 398 Images	Accuracy: 90%	2022	[171]
Organ Segmentation	Mask R-CNN	In-house	55 Videos 8 Hospitals 1578 Images	Dice Score: 97.65%	2022	[270]
Smoke Removal	CycleGAN	In-house	10 Videos 6000 Images	Accuracy: 93%	2022	[276]
Organ Segmentation	U-Net EfficientNet-b5	In-house	20 Subjects 506 Images	mDSC: 0.90	2022	[271]
Gaze Detection & Segmentation	YOLOv3 U-Net InceptionV3 MobileNetV2 ResNet-50	In-house	4003 Videos	Accuracy: 77.68, 75.67, 90.16 IoU: 0.85	2022	[277]

This section illustrates the datasets employed by the articles encompassed in this study. Furthermore, around 35 datasets are adopted by the surveyed articles, however, for the purpose of concision, only 10 datasets are described. The Table 3.12 contains the information of publicly available datasets including name, year of publication, modalities, and a short description. According to this study, the JIGSAWS is the most employed dataset in the RAS that encompasses video and kinematic data (Figure 3.5).

3.4.1 JIGSAWS

The well-known gesture and skill assessment dataset JIGSAWS comprises of kinematic and video data for suturing, needle passing and knot-tying tasks. This dataset was recorded by 8 experts on da Vinci surgical system performing the five repetitions of suturing, knot-tying and needle passing procedures [93]. The Figure 3.5 shows that the JIGSAWS is most extensively applied dataset by the studies included in this case study, however, not all the authors have utilized the complete dataset but a part of the dataset. The JIGSAWS dataset contains 163 videos and kinematic data. The brief description of the dataset is provided in the Table 3.12.

3.4.2 MICCAI Datasets

The Medical Image Computing and Computer Assisted Intervention (MICCAI), officially known as *The MICCAI Society* was established in July 2004 [94] as a non-profit organisation. *The MICCAI Society* holds the annual competitions with the purpose to bring the researchers, clinicians, and engineers together to advance in the field of biomedical engineering. The MICCAI offers a wide range of datasets each year with different objectives and holds the competitions. A worth considering number of articles take advantage of datasets provided by challenges for biomedical analysis.

The most common challenge of MICCAI is EndoVis challenge, which is organised every year since 2015. The sub-challenges of EndoVis contain datasets for wide variety of tasks including instrument detection, segmentation, boundary detection, workflow analysis, skill assessment, etc. Further information about the MICCAI and its EndoVis challenge can be found in the Table 3.12.

3.4.3 Cholec80

The Cholec80 is another famous cholecystectomy dataset. It comprises of 80 cholecystectomy surgery videos recorded during the surgical interventions performed by 13 experts at 25 Frames Per Second (FPS) [95]. The Cholec80 data is widely used for tool presence detection and phase recognition as evident by the Figure 3.5 and Table 3.12. The original data contains the videos where the tool cannot be easily visualised by naked eye which makes the detection and segmentation a challenging task.

The Cholec80 dataset is further split into two groups of 40 videos each, namely for training and testing. Several different versions of Cholec80 are also available in literature.

3.4.4 ATLAS Dione

The ATLAS Dione dataset contains the video data of ten surgeons having different expertise level and record six different surgical procedures on da Vinci surgical system at Roswell Park Cancer Institute, Buffalo, New York, USA [96]. The surgical procedures include basic robotic surgery task to high-level surgical processes. The motivation behind this data generation was unavailability of annotations in JIGSAWS dataset. The videos in the dataset are annotated in the frames of 854×480 pixels each and the annotations are provided in XML format. Further details about the ATLAS Dione dataset can be found in the Table 3.12.

3.4.5 UCL dVRK Dataset

The UCL da Vinci Research Kit (dVRK) comprises 14 videos of 300 frames each of segmentation task. It also contains six videos of robotic kinematic recorded at 300 frames each. All the recorded frames are 720x576 pixels. The frames contained the camera artefacts which are later cropped at 720x576 pixel resolution. The dataset is recorded using four consecutive steps that are repeated to record the video and kinematic data. For the further information on the setup and the data acquisition methods, interested reads are referred to the [176].

3.4.6 M2CAI16 Dataset

The M2CAI16 challenge is a satellite event of MICCAI offering two different datasets including workflow and tool detection, which are enlisted hereunder as well as in the Table 3.12.

3.4.6.1 M2CAI16-Tools Dataset

This dataset was generated with the collaboration of University Hospital of Strasbourg and the Hospital Klinikum Rechts der Isar in Munich, Germany [96]. The dataset contains 41 laparoscopic procedural videos of the cholecystectomy with eight distinct phases. Out of the total 41 videos, 27 and 14 videos are for training and testing purposes, respectively.

Table 3.12 The summary of the publicly available datasets used by the articles included in this study

Name	Data	Data Description	Procedure	Purpose	Year	Ref
MICCAI	Video	41 Videos 25 fps 1920 × 1080 Resolution	Subject to Change	Detection Recognition Segmentation	2004	[94]
JIGSAWS	Video Kinematic Data	103 Videos 39 Sequences 8 Users 10 Gesture Classes	Suturing Knot Tying Needle Passing	Gesture Recognition Skill Assessment	2014	[93]
Cholec80	Video	80 Videos 13 Surgeons 25 fps 854 × 480 Resolution	Cholecystectomy Surgery	Phase Recognition Tool Detection	2016	[278]
M2CAI16 Tool	Video	15 Procedures 2532 Frames	Cholecystectomy Surgery	Tool Detection	2016	[279]
M2CAI16 Workflow	Video	41 Videos	Cholecystectomy Surgery	Workflow Analysis	2016	[280]
M2CAI16 Location	Video	3141 Annotations 7 Instrument 10 Videos	Cholecystectomy Surgery	Tool Detection Skill Assessment	2016	[279]
ATLAS Dione	Video	10 Surgeons 6 Tasks on daVinci 99 Videos 854 × 480 Pixels 22467 Images	6 Different Surgeries	Tool Detection Action Recognition	2017	[96]
DeepLision	CT Images	32000 Images 4400 Patients	Whole Body	Lesion Detection Semantic Segmentation	2018	[281]
Lapro425	Video	425 Videos 9 Surgeries	Laparoscopy	Flow Analysis	2019	[246]
UCL dVRK	Video Kinematic Data	14+6 Videos 300 Frames Each 720 × 576 Pixels	Different Surgeries	Tool Segmentation	2020	[176]
CholecT40	Video	40 Videos 13 Surgeons 25 fps 854 × 480 Resolution	Cholecystectomy Surgery	Action Recognition	2020	[248]
Sinus Surgery-C	Image & Video	10 Videos 5-23 Minute 320 × 240 Resolution 30 fps	Sinus Endoscopy	Smoke Detection Tool Shadow Instrument Segmentation	2020	[282]
FlapNet	Video	62 Minute Videos	Lobectomy	Tissue Segmentation Tool Segmentation	2020	[210]

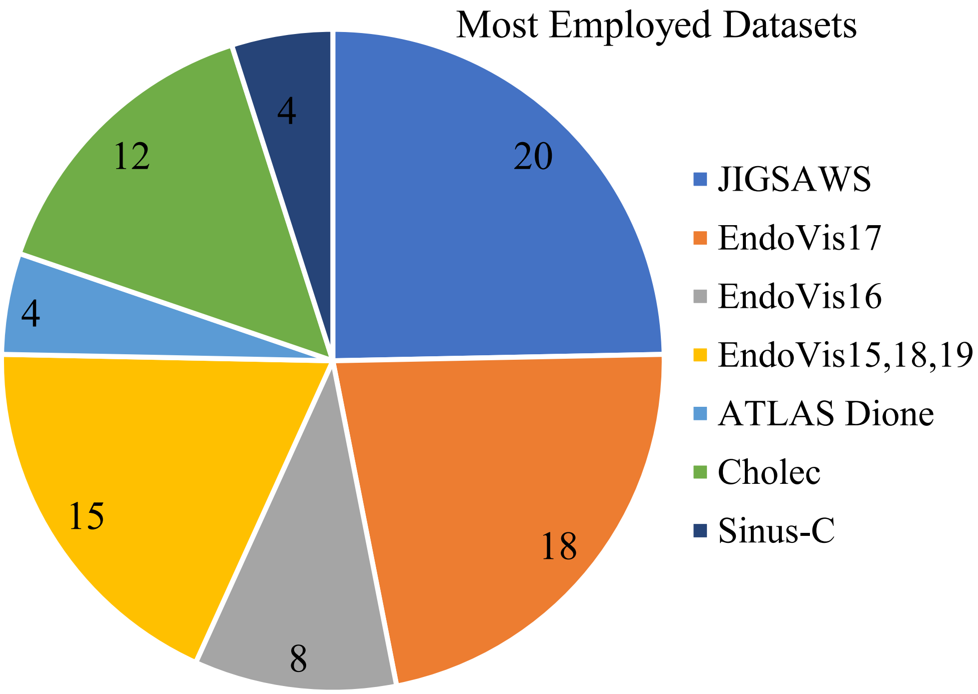


Fig. 3.5 Frequency of the used datasets by the number of studies. Articles that adopted two or more datasets are counted respectively.

3.4.6.2 M2CAI16-Workflow Dataset

The M2CAI16-Workflow dataset was generated at University Hospital of Strasbourg [278]. Similar to the previous data, this dataset also contains laparoscopic videos of cholecystectomy. A total of 15 procedures were recorded, ten and five for training and testing purposes, respectively.

3.4.6.3 M2CAI16-Location Dataset

The M2CAI16-Location dataset is extension of the aforementioned M2CAI16-Tool dataset [279]. It additionally contains the annotations of the tools and locations.

3.4.7 Laparo425

The Laparo425 dataset contains the laparoscopy videos of nine distinctive classes. The dataset was recorded at the University Hospital of Strasbourg. As its name suggests, it contain 425 procedures recorded at 25 FPS and down-sampled at one FPS for experiments.

3.4.8 CholecT40

The CholecT40 datasets comes from the Cholec80 dataset. It contains 40 videos from the Cholec80 dataset and annotates them to action triplets. The 128 different action triplets are introduced and recognised using video data [248].

3.4.9 Sinus-Surgery-C Dataset

This dataset contains the endoscopic sinus surgery images which are annotated for surgical tool segmentation task. The dataset comes from the BioRobotics Lab at the University of Washington, USA [282]. The segmentation of the dataset is not easy since it contains smoke and shadows of the instruments. As soon as the movement occurs, the blue images generate that make the dataset more challenging for instrument segmentation tasks.

3.4.10 DeepLesion

The DeepLesion dataset is large dataset of National Institute of Health of USA that contains 32000 lesions of the CT images. The total number of 4400 unique patients were involved during the data generation [281]. This dataset offers a great diversity because of its diverse collection of images from liver, lungs, lymph, and many more human body organs. The further details about all datasets are also provided in the Table 3.12.

3.5 Risks and Challenges

This section describes the potential pitfalls that hinder the development of autonomous robotic surgical systems. The first part of the section illustrates the technical and technological challenges that are either under development phases or soon will be. However, the second part demonstrates the legal and ethical concerns of the DL guided computer-aided interventions.

A big thanks to the advancement in the AI, the healthcare industry has not only improved the diagnostic methods but also moved to surgical robots. However, the other side of the DL in medical domain is still darker. The technical risks include the design of the robotic components, the precision in complex scenarios, and unpredictability of the amount of surgical procedure assignment [283] among numerous others. The success rate of surgical systems is increasing day by day [283], nevertheless the high associated costs are not negligible [284]. These costs can be traced back to component design and continue through the operating room to the maintenance expenditures.

Additionally, the surgical systems are designed with limited allowed movement and the dexterity, which have a positive and a negative side. The controlled movement will fail if any unscheduled task originates during the surgery because of no or less adaptability expertise [285]. In the meantime, a robotic system can physically harm the patient and damage the involved components. The safety of the patient under the surgery is at risk, not only by hardware part of the robotic system but also by the software [286].

In a semi-autonomous surgical system, a minute human error can put the life of patient in danger [287]. The training and the testing capabilities of human are not as precise as machine, however completely relying on machine with present day technology could produce less efficient outcomes in general [288, 289]. Therefore, all these concerns require detailed studies and considerations.

Similarly, all the perks DL offers are appreciable, however, there are few things beyond technology. The legal and ethical issues concern the privacy of the medical data of the patients [290–292]. Another potential threat is cybercrime involving the genetic information of the patient and potential hack of the robotic system [293]. The prospective cyber attacks to seize the control of surgical system can lead to devastating results including lethal physical harm to patients, as discussed by Bonaci et al. [294]. The software designer and the developers should come forward and certify that the delivered products are not vulnerable to attacks. An automatic rescue service should be activated in cases of emergency i.e. power cuts, jamming, transmission, etc. Unlike autonomous driving, there exist no legal standard methods to define the level of degree to which a robotic system can be autonomous [295]. The National Health Service (NHS) of the United Kingdom came under similar attack back in 2017 which affected medical devices nationwide [296]. Moreover, the transparency is another worrisome issue because of the blackbox decision making nature of the DL models. The recent advent of eXplainable AI has taken care of transparency and interpretability problems, however, the field is growing and requires further exploration. Finally, all the aforementioned applications are accelerating the dread of physical harm by the AI systems and eventually reduce the human involvement. The additional intervention of government and giant companies spreads fear among the undergoing patient that demands and requires a code of conduct.

The General Data Protection Regulation (GDPR) by European Union also states the concise and transparent information provision and privacy protection of users [42]. This ensures the maximal control, however the full enforcement of the law across the board is yet to be experienced.

Furthermore, the fully autonomous robotic surgical system should first go through several security checks to answer the concerns of physicians, engineers, patients and general public. Who will be responsible in case of harm to the undergoing patient or patient's data? Is it the robot? Or the engineer developing the robot? Or anyone else? Successfully building the autonomous robotic systems will leave nothing to human but the responsibility which brings culpability and liability. The author, in no case, is against the development of fully autonomous robots, rather demands a proper regulation before the train is missed.

Conclusively, the cyber-security of the robotics and particularly the surgical robotics is big market where researcher need to explore because a medical robotic system can not be judged based on average results but best possible outcomes are indispensable (better than human), otherwise, a medical robot looses the whole point to be developed for surgical applications.

3.6 Discussion

During the recent decade, the DL models have made substantial impact on healthcare domain (e.g. CAD systems). The large scale availability of surgical datasets and straightforward data acquisition protocols encourage cross-domain research synergies in order to reach fully autonomous robotic surgery. Recently, numerous DL models have been applied on medical images to capture the relevant information and provide to the RAS system for surgical procedures (Figure 3.1).

This case study analyses and summarises the contribution of the DL architectures in the field of image-driven surgical robots. The analysis reveals the current interests of researchers in image-guided DL based RAS is increasing over time. Inside the RAS, the majority of articles contribute to the tool detection and segmentation tasks. One reason behind detection and segmentation task selection is the serene accessibility to enormous surgical tool data. The other reason can be the working mechanism of DL models. Since the DL models learn from patterns, the tools segmentation tasks are arduous because of the presence of smoke, reflection of instruments, and vessels in the surgical dataset.

During this study, the author found that the CNN is most commonly employed DL method because of its successful history with image data. The Figure 3.4 and Table II through Table XI are evident of the aforementioned claim. Not only the CNN is most applied method, but also it outperforms other techniques. However, as depicted in the Table 3.1, the combination of several different models yields improved results. Most of the articles dealing with action recognition applications (e.g. surgical gesture recognition and segmentation, trajectory segmentation, phase and state estimation, etc.) consider the combination of traditional static CNN architectures along with the RNN/LSTM models. The combination of CNN and RNN lies in the fact that CNN are able to automatically extract spatial features within images, whereas RNN have been designed to capture the temporal information. However, it is worth noting that not all the articles related to action segmentation employ a RNN. In fact, some authors choose 3D CNNs that are able to simultaneously capture both spatial and temporal features of the action/motion behaviour. These kind of solutions can directly extract information from multiple video frames through 3D convolutions. Comparing the two different approaches it is not easy to determine which approach is the best, a future work might investigate deeper this aspect. However, it can surely stated that the main limitation of 3D CNN is related to the increased dimension due to 3D convolutions, thus a bigger amount of data and time is needed for training such models.

Similarly, the most widely used dataset is JIGSAWS that contains video and kinematic data (Figure 3.5 and Table 3.12). Authors proved that the fusion of video data into the kinematic data leads to better performance outcomes. The annual competition organised by MICCAI for several varying imaging tasks also brings new datasets. Although huge number of studies employ in-house datasets, a total number of 35 different publicly available datasets are used either partially or fully.

This fact reveals that the amount of available data is significant and has attracted scientists and boosted the research. Another benefit of extensive data availability is the possibility of comparison between the novel conducted studies.

A considerable number of papers have taken benefit of pretrained networks, as can be seen through the Table 3.1 and Table 3.10. These models are heavily trained on ImageNet dataset [251]. The biggest reason of using pretrained nets is the unavailability of sufficient amount of data to train a DL architecture from the scratch. In the surgical workflow analysis task, numerous authors adopted in-house datasets which led them to use pretrained models. Since these models are pretrained, therefore, they come with the added benefit of lower time consumption. The most commonly used pretrained net remained VGG architecture followed by the ResNet and MobileNet.

With the increasing development of DL models, and the massive success with image modalities, the research in the field of image-guided robotic surgery is inevitably growing. Firstly, image data acquisition methods are smooth with least harm (tolerable) and do not require colossal acquisition cost. Secondly, image data contains bulk of information about the patient and helps understand internal state of patient without incision or physical injection.

Moreover, analysis and preprocessing on the image modalities is relatively easy and also, the DL methods work better with the image data. Therefore, the future of MIS, RAS and computer-aided interventions relies mainly on the image and video data, and of course DL.

Finally, based on the exhaustive analysis, the study points out to the two major concerns, first for the medical experts and the other for DL engineers. The recorded image modalities and videos contain smoke, instrument reflections, and surgeon's movements (i.e. hands, reflections), therefore, the DL algorithms learn from the unnecessary features which may bias the results. Secondly, not only the DL methods require huge amount of time for training purposes that needs to be optimized, but also DL is blackbox in decision making mechanism which concerns the clinicians and the patients.

3.7 Summary

In this study, technical articles concerning image driven computer-assisted interventions incorporating DL models are selected for the case study. The intensive text assessment indicated that the selected 184 articles can be grouped into four categories including: 1) Surgical Tools, 2) Surgical Processes, 3) Surgical Surveillance, and 4) Surgical Performance/Assessment. The key findings include: a) Surgical Tools is most studied topic which comprises Surgical Tool Detection and Surgical Tool Segmentation (45% of total articles), b) CNN is most widely applied DL topology (roughly 54% of total articles), c) the gesture recognition articles incorporate JIGSAWS dataset (around 77% of articles in relevant subcategory), whereas MICCAI datasets are top consideration for detection and segmentation tasks (around 60% of articles in relevant subcategory), d) VGG remains the widely accepted pretrained network especially when available dataset was not large enough, f) the most studied applications appear to be cholecystectomy and prostatectomy, g) for gesture and trajectory applications, suturing task is frequently studied application area, h) the fusion of kinematic data with image data yields better results. Considering the characteristics of the proposed case study, the author believes that the main limitation of the study concerns the lack of deep details of the pre-processing and processing approaches proposed in the reviewed papers to solve the problems related to each application category. However, this was not the original scope of the case study, since providing the general overview of the topics under discussion required gigantic efforts. In the context of future direction, the development of fully autonomous RAS system appears highly promising and fascinating research topic. Additionally, self-supervised learning based models can greatly improve the environment in operating room. Finally, a steady walk from the weak AI to strong AI and to the super AI can lead to the notable breakthroughs.

Chapter 4

Deep Learning Driven Fusion Biopsy for Prostate Morphology

4.1 Deep Learning framework for Prostate Segmentation

In this section of the work, the accurate, reliable, fast, and hence semiautomatic approach for prostate segmentation from TRUS images has been proposed. The pipeline can be exploited without having acquired a specific dataset for the transducer in consideration. The approach, like other existing works, is based on the theory of deformable superellipses. Two kinds of methodologies can be exploited for achieving segmentation with superellipses. In the first, image characteristics, such as edge maps or region energy, are employed for performing automatic image segmentation. In the second case, the geometry of the prostate is inferred from user-defined points.

Though extensive experiments have been carried out to outline the best guidelines that a human operator should take into consideration when using the proposed algorithm for performing a procedure in order to minimize the number of points required and maximize

Table 4.1 *Prostate gland datasets description.* As imaging modalities, only ZENODO contains TRUS images, even in a very limited quantity, being only 3. Fiducial points are also present only in the last dataset. Seg stands for Segmentation, Reg for Registration.

Dataset	**Imaging modality**	**Task**	**Number of images**	**Ground Truth segmentation**	**Fiducial points**	**File format**
PROMISE12 [297]	MR (T2W)	MR Seg	50	✓	✗	NIfTI
SAML [298, 299]	MR (T2W)	MR Seg	116	✓	✗	NIfTI
ZENODO [300]	MR (T2W), TRUS	TRUS Seg, MR/TRUS Reg	3, 3	✓	✓	NRRD

the segmentation accuracy. In any case, carrying out a second iteration can mitigate eventual problems that arise after the suboptimal placing of points in the first iteration.

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

4.2 Introduction and Background

Prostate cancer is a major health problem and represents the most common cancer in the male population, accounting for 18.5% of all the cancers diagnosed in humans [301]. The number of new cases worldwide crossed 1,275,000 and caused approximately 360,000 deaths merely in year 2018 (3.8% of all deaths caused by cancer in men) [302].

Numerous imaging modalities are exploited for prostate cancer diagnosis, treatment and follow-up. The TRUS, MRI, and CT are the most common employed imaging modalities [303]. Each technique provides different information and is used for several divergent clinical scopes. During biopsy procedures, TRUS is commonly employed since it is an inexpensive, portable and real-time methodology [304]. MRI is mainly adopted for diagnosis and treatment planning [305]. In fact, this modality has a better soft tissue contrast and allows a more efficient lesion detection and staging in patients affected by prostate cancer.

As can be seen from Figure 4.1, the TRUS images suffer from problems as speckle, low contrast and shadow artifacts [306]. Calcification and acoustic shadowing make the automatic segmentation of prostate region a very complex task [307]. The prostate usually appears like a hypoechoic mass encompassed by a hyperechoic region [308]. CT scans are useful in determining if prostate cancer has spread to bone tissues or to assess the effectiveness of the brachytherapy [309].

In the clinical practice, majority of prostate cancer cases are diagnosed, prior to symptoms development, thanks to the Prostate-Specific Antigen (PSA) [310] levels in the blood and rectal examination. In order to achieve more satisfactory information, MRI is the election modality, with PI-RADS v2.1 being the standard for finding interpretation [311].

The standard prostate biopsy involves the extraction of 10 to 12 tissue samples. Since there is no guarantee that sampling prostate in these regions is the most effective way to obtain the regions with cancerous tissue, fusion guided prostate biopsy is becoming now the preferred modality for most urologists and surgeons. In this way, suspicious areas found in the MRI of the prostate can be targeted during the prostate procedure exploiting the fusion with the real-time TRUS imaging, also allowing a better view of the biopsy needle.

Advantages of this approach comprise of the following: more accurate sampling of the cancerous tissue in the prostate gland; less amount of patient tissue is extracted; less pain and less risks for the patient, including faster recovery time [312].

In order to implement a fusion prostate biopsy framework, segmentation of prostate gland must be obtained from both TRUS and MRI domains. Exceptions involve systems in which images are manually registered by the user at procedure time, by superimposing MRI over the TRUS.

Since MRI is acquired days before the prostate biopsy, it is not fundamental that its segmentation is performed real time. Even though, manual segmentation of prostate from MRI is a tedious task and prone to inter- and intra-radiologist variation [297], so the exploitation of an automatic method grounded on the nnU-Net framework [313] for this task can further ease the procedure and improve its diagnostic accuracy.

It is worth noting that nnU-Net does not denote a novel network topology, loss function, or training procedure. Indeed, nnU-Net stands for "no new net". The strength of the nnU-Net framework comes from the systematization of all the steps which were usually manually tuned in the training pipeline of semantic segmentation architectures, including data augmentation, hyperparameters' tuning, test-time augmentation, and ensembling.

Instead, manual segmentation of the prostate gland from TRUS has to be realized real-time during the prostate procedure, therefore the need for a fast and effective methodology for this task is really fundamental in the clinical practice.

Ghose et al. performed a comprehensive survey which focused on methods for prostate segmentation in TRUS, MR, and CT images [303]. Prostate gland segmentation eases multi-modal image fusion for tumor localization in biopsy. Manual annotation of radiological images is a tedious and error prone task, which also has problems like inter- and intra-radiologist variability. Fully automatic methods, as those based on DL, require huge annotated data, usually of the same transducer that will be used for the procedures, since there is a high variability in ultrasound image quality across vendors. Nonetheless, when data is available, DL methodologies show their strength, as is the case for Deep Attentional Features (DAF) [314] and DAF 3D [26]. The shortcoming of these techniques is that they cannot be applied before having acquired a dataset with images of the same ultrasound device that will be adopted during procedures.

Mahdavi et al. [315] proposed a semi-automatic prostate segmentation method that can be applied in prostate brachytherapy setups. The 3D geometric model of the prostate is created based on prior knowledge of the shape of the gland and on the assumption that the prostate has a tapered ellipsoidal shape and is slightly warped posteriorly due to the presence of the

TRUS probe. They used, as prior shape of prostate gland, a tapered and warped ellipsoid. The proposed segmentation algorithm requires a manual initialization of the physician: on the mid-gland image the user selects six boundary points following a specific criterion. The main disadvantage of this method is that it requires the user to put initialization points in a very precise way, and relies deeply on these points, posing problems if they are slightly inaccurately placed or when the prostate region has an irregular shape.

Gong et al. [316] incorporated deformable superellipses in a Bayesian segmentation framework, exploiting an edge detection algorithm for discovering prostate boundaries. They show the capacity of deformable superellipses to capture the prostate shape in various anatomical zones. The main limitation of this method is that it requires to have an initial contour that is similar to the real boundaries of the prostate gland. To overcome this issue, Saroul et al. [317] proposed a variational approach, exploiting the implicit representation of a superellipse for modeling the active contour.

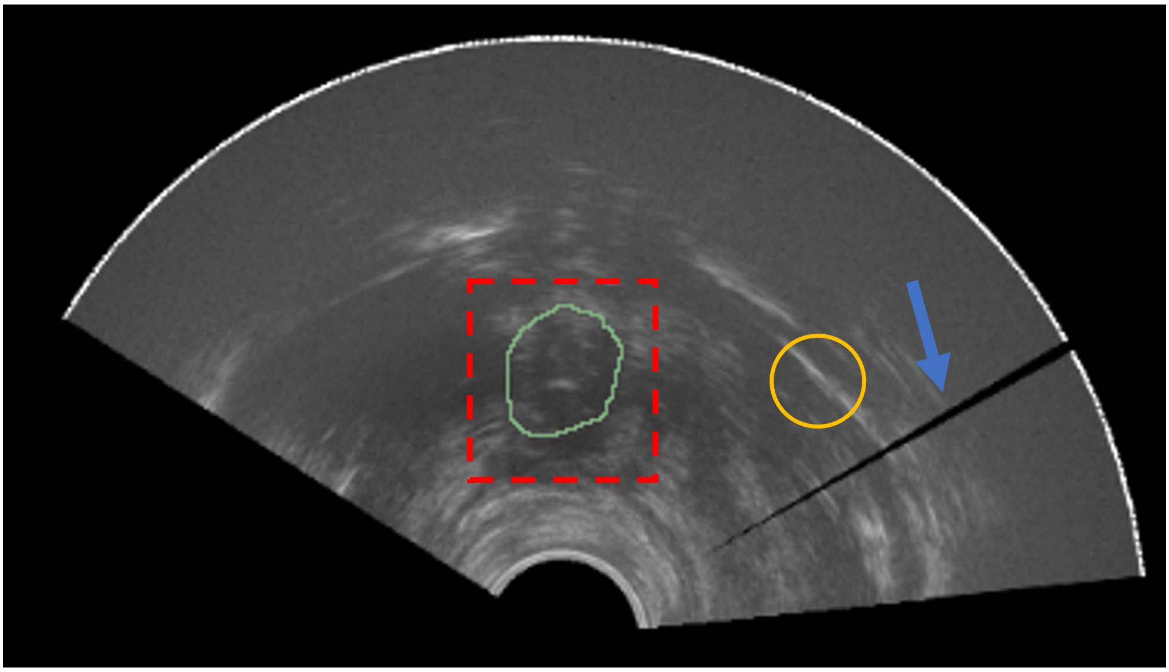


Fig. 4.1 *Prostate gland in TRUS image*. Prostate apex (ground truth mask in green) is not well distinguishable from the rest of the image (red dashed box). Yellow circle represents an example of region with low signal-to-noise ratio. Blue arrow denotes a shadow artifact.

4.3 Materials

Fedorov et al. made a publicly available dataset containing anonymized imaging data of the human prostate of N=3 patients [300]. This dataset will be referred to as ZENODO throughout this section. For each patient, both MRI and TRUS examinations have been acquired. The former serves the purpose of staging the disease and the latter of allowing volumetric examination for preparing brachytherapy implant. Both modalities are 3D scalar images. Annotations provided by Fedorov et al. include manual segmentation of the whole prostate gland for both MRI and TRUS, and fiducial points placed in specific anatomical sites to improve subsequent image registration and fusion. In detail, fiducials are placed at urethra entry into the prostate at the base (UB), verumontanum (VM), and urethra entry into the prostate at the apex (UA).

In order to validate DL models for the task of MRI prostate segmentation, also the datasets PROMISE12 [297] and SAML [298, 299] have been included in the analysis.

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

Sample images for both domains, TRUS and MRI, are reported in Figure 4.2. A summarized table for the considered materials is provided in Table 4.1.

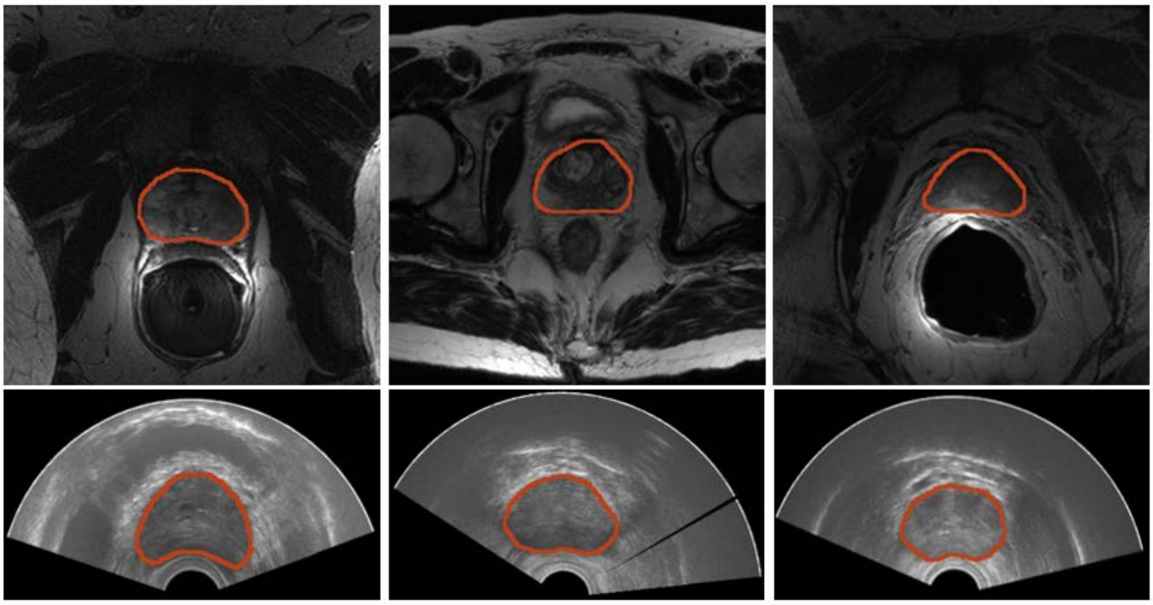


Fig. 4.2 *Samples of images from MRI and TRUS modalities.* (Top) Prostate MRI. From left to right, a sample image for each of the datasets PROMISE12, SAML, and ZENODO is shown. (Bottom) Three sample prostate TRUS from the ZENODO dataset.

4.3.1 Workflow

The workflow employed for achieving image fusion, starting with segmentation for both imaging modalities, namely TRUS and MRI, is reported in Figure 4.3. In clinical practice, segmentation does not happen at the same time, since MRI segmentation can be achieved preoperatively, whereas TRUS segmentation has to be obtained intraoperatively, at the start of the prostate biopsy procedure.

Segmentation from MRI involves a pre-processing stage so that images can be fed to a deep learning architecture, the nnU-Net. Lastly, post-processing is performed, with the aim to remove noisy elements from images (e.g., only one connected component is expected), increasing segmentation accuracy. The described operations can be carried out in a fully automatic way. MR images are especially important for identifying the target region for biopsy since they have better contrast than other imaging modalities. Details are described in Section 4.3.2.

Segmentation from TRUS is achieved with a semiautomatic algorithm, which requires input points from the user. The physician has to annotate points in at least three slices of the prostate gland in axial planes, taking care when placing points in the deformed zones (the transducer itself introduces deformation). Starting from this point, a deformable superellipse is fitted with an optimization algorithm. Then, interpolation is employed to achieve the 3D reconstruction of the prostate gland. The entire procedure is explained in Section 4.3.3.

Then, with both segmentation masks from TRUS and MRI modalities available, registration can be performed, enabling image fusion, which allows tissue coming from both modalities to be seen at the same time. Optionally, a set of anatomical landmarks can be inserted by the user to ease and constrain the registration optimization step. The procedure is presented in Section 4.3.4.

4.3.2 MRI Segmentation

The semantic segmentation of the prostate gland from MRI can be efficiently met via DL techniques, as fully convolutional neural networks [318]. Semantic segmentation, which poses the basis for subsequent classification and characterization tasks [44, 319], is essential in numerous clinical applications including artificial intelligence in diagnostic support systems, therapy planning, intraoperative assistance, and monitoring of tumor growth.

As introduced in Section 2.1, semantic segmentation is a Computer Vision task that can be computed with DL algorithms and consists in labeling each pixel of an input image, without recognizing the different instances of objects [320, 321]; it is possible to see semantic

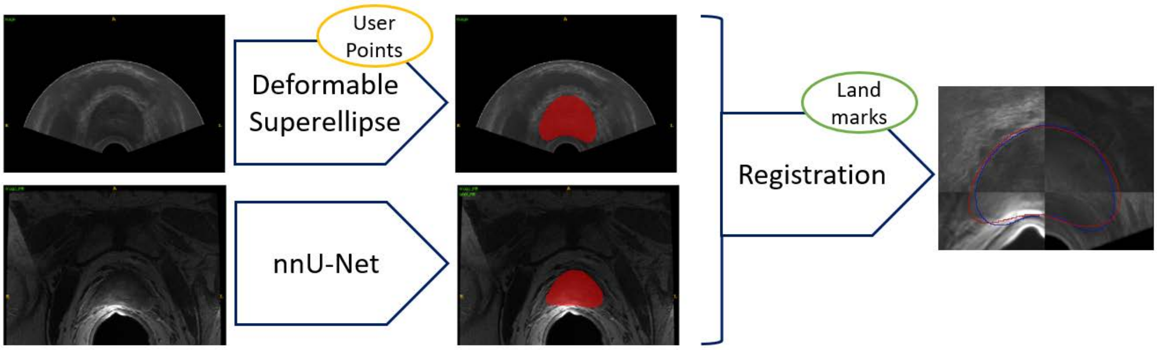


Fig. 4.3 *Workflow for TRUS and MRI segmentation and subsequent image fusion.* Segmentation from TRUS is achieved in a semiautomatic way by fitting a 3D model based on deformable superellipses starting from user-defined points in at least three slices. Segmentation from MRI has been performed fully automatically by exploiting the nnU-Net framework. Registration can be either performed in an automatic way, or the user can add anatomical landmarks to constrain the space of transformations.

segmentation as a problem of conversion from image to image, where the input image is the original image and each pixel intensity value of the output image indicates the relation of that pixel to the associated class [322].

Most semantic segmentation architectures are based on encoder-decoder networks. The process of feature extraction or sub-sampling is carried out by the encoder. Decoding is an up-sampling operation, in which the spatial information output from the encoding layer is reconstructed, increasing the spatial resolution. The encoder-decoder structures have been implemented in different convolutional network architectures, including SegNet [323], U-Net [324], U-Net 3D [325] and V-Net [326]. Besides prostate segmentation, applications in medical imaging tasks of those architectures encompass liver vessels delineation [327], segments classification [328], lung COVID-19 lesions segmentation [329], and vertebrae segmentation [330].

In the work presented in this section, to perform the semantic segmentation of the prostate gland from MRI, the nnU-Net framework has been exploited. In this way, semantic segmentation tasks can be tackled with standardized pipelines [313, 331]. The employed architecture is based on those of U-Net and U-Net 3D. The nnU-Net framework was originally conceived during the Medical Decathlon Segmentation Challenge [332], where it emerged as the leading approach in all tasks. The advantages of this method consist of automatic configurations of pre-processing, data augmentation, training, inference, and post-processing. Parameters to set for training nnU-Net include the number of epochs, initial learning rate, batch size, patch size, and the combination of dice loss and cross-entropy to implement as the model loss function.

4.3.3 TRUS Segmentation

Segmentation of anatomical structures in noisy data, such as TRUS images, is a complex task since boundaries are not clearly defined, as shown in Figure 4.1. Therefore, the adoption of prior information about the geometric structure of interest is useful to constrain the model deformation [333, 334]. Deformable models can be used to achieve this result.

Geometry, physics, and mathematical optimization lie the foundation for the segmentation algorithms based on deformable models [333]. The constraint on the model shape is derived from geometry, the evolution of the shape in space is governed by physical theories, and the operation of fitting the model to the accessible data is made possible by optimization theory [335]. Segmentation of anatomical structures in deformable models is achieved by exploiting an energy minimization framework. Two kinds of mathematical terms are considered: internal and external energies. The deformable model is propagated in the direction of the object contours by external energies, whereas the smoothness of the boundaries is preserved by internal energies.

The deformable model framework includes various methodologies, such as deformable mesh, active shape models, level sets, active contour models, and curve fitting [303]. More advanced approaches may include a mixture of these techniques, with the idea that merging information concerning boundaries known *a priori*, region, shape, and features of the prostate region can provide more accurate models, like the deformable superellipse formulation of Gong et al. [316].

4.3.3.1 Shape Models

In a wide variety of medical imaging scenarios, the general location, orientation, and shape of the objects of interest are known *a priori*. As reported by previous studies concerning TRUS images, prostate contours appear smooth and with a closed-near convex shape [316]. This information can be embedded into the deformable model in different forms: initial conditions, way of constraining model shape parameters, and the procedure for model fitting. Global shape properties can be modeled with parametric shape models. The advantage of this technique is not requiring the presence of anatomical landmarks.

Furthermore, representation of the shapes can be tackled with many different methods [336, 337]. For instance, Tutar et al. [338] proposed to model the 3D prostate boundaries with spherical harmonics of degree eight. Local deformations can be controlled, thanks to the exploitation of parameters, leading to the capacity of modeling complex shapes. On the other side, there is an increment in the computational complexity.

Similarly, reducing the number and range of parameters can allow modeling the global shape in approaches that are stable and rapid from a numerical perspective, leading to more compact representations. In the following section, the deformable superellipse, a powerful model for the geometry of the prostate gland [316], is introduced. When the deformable superellipse is not capable of properly capturing all the nuances of the prostate region in a 2D slice, bidimensional B-splines [339] can be exploited in the proposed approach, obtaining very refined results but not losing the possibility to model a regular 3D shape with a relatively low number of parameters.

4.3.3.2 Deformable Superellipses

Superellipses consent to obtain a natural generalization of the ellipses' shapes. Different base geometrical shapes can be modeled through superellipses, such as ellipses, parallelograms, rectangles, and pinched diamonds by handling a small number of parameters [340, 341]. Examples of shapes that can be modeled by superellipses are portrayed in Figure 4.4. The straightforward 3D generalization of the superellipse, the superellipsoid, has not been considered since it makes assumptions about the 3D regularity of the prostate shape which are too simplistic.

A centered superellipse can be expressed in the following **parametric** form, as reported in Equation (4.1):

$$\begin{cases} x = a_x \cdot |cos(\theta)|^{\frac{2}{\varepsilon}} \cdot sign(cos(\theta)) \\ y = a_y \cdot |sin(\theta)|^{\frac{2}{\varepsilon}} \cdot sign(sin(\theta)) \end{cases} \tag{4.1}$$

where the size parameters $a_x > 0$, $a_y > 0$ define the length of the semi axes, and $\varepsilon > 0$ specifies the squareness in 2D plane, as shown in Figure 4.4. The corresponding **implicit** form is given by the Equation (4.2):

$$\left|\frac{x}{a_x}\right|^{\varepsilon} + \left|\frac{y}{a_y}\right|^{\varepsilon} = 1 \tag{4.2}$$

The *inside-outside* function is reported in Equation (4.3):

$$f(x,y) = \left|\frac{x}{a_x}\right|^{\varepsilon} + \left|\frac{y}{a_y}\right|^{\varepsilon} \tag{4.3}$$

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

The superellipse model does not permit, in this version, molding all deformations which are required to build a proper representation of the prostate gland. Nonetheless, geometric deformations, such as translation, rotation, tapering, and bending, can result in a broad range of shapes that are modeled by the deformable superellipse [342, 343]. Moreover, these transformations can be modeled with a few parameters, given that translation with respect to an axis, rotation, tapering, and bending are described each with a single parameter. Deformable superellipse can then be characterized by a parameter vector **p**, defined as in Equation (4.4):

$$\mathbf{p} = \{a_x, a_y, l_x, l_y, r, \varepsilon, t, b\} \tag{4.4}$$

where ε is the squareness parameter and a_x, a_y are the semi-axes lengths defined above. Other parameters are those involved in the global similarity transformations for superellipses [316]: l_x, l_y are the translations along x and y axes, r is the rotation angle, t and b model the tapering and circular bending on the y-axis, respectively.

Details of all these geometric transformations are reported in Paragraph 4.3.3.2-**Geometric Transformations**, whereas inverse transformations are reported in Paragraph 4.3.3.2-**Inverse Transformations**. Examples of deformable superellipse modeled by variations in tapering and bending are reported in Figure 4.4.

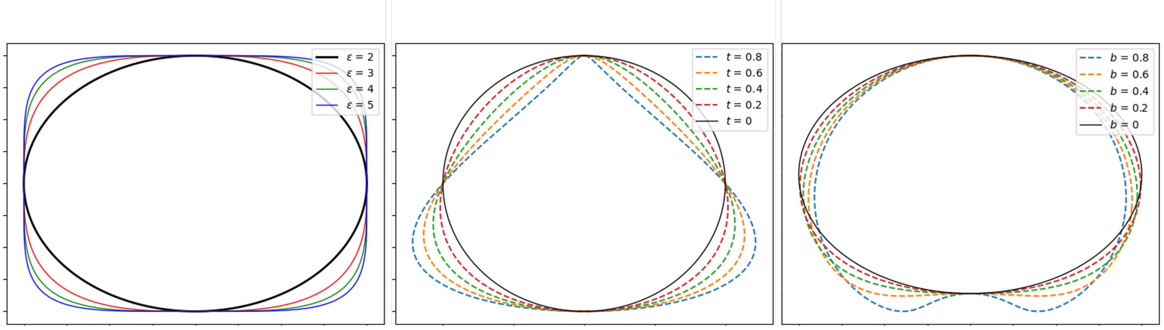


Fig. 4.4 *Deformable superellipse modeling capabilities examples*. The left image represents the superellipse varying squareness ε parameter, whereas the middle one depicts the deformable superellipse varying tapering t parameter, and the right one portrays the deformable superellipse varying circular bending b parameter.

Geometric Transformations Translation (l_x, l_y):

$$\begin{cases} x' = x + l_x \\ y' = y + l_y \end{cases} \tag{4.5}$$

Rotation (r):

$$\begin{cases} x' = x \cdot cos(r) - y \cdot sin(r) \\ y' = x \cdot sin(r) + y \cdot cos(r) \end{cases} \tag{4.6}$$

Linear tapering along y-axis (t):

$$\begin{cases} x' = x \cdot \left(\frac{t \cdot y}{a_y} + 1\right) \\ y' = y \end{cases} \tag{4.7}$$

Circular bending along the y-axis (b):

$$\begin{cases} x' = (\frac{a_y}{b} - y) \cdot sin\left(\frac{x}{\frac{a_y}{b} - y}\right) \\ y' = \frac{a_y}{b} - (\frac{a_y}{b} - y) \cdot cos\left(\frac{x}{\frac{a_y}{b} - y}\right) \end{cases} \tag{4.8}$$

Inverse Transformations Inverse translation (l_x, l_y):

$$\begin{cases} x = x' - l_x \\ y = y' - l_y \end{cases} \tag{4.9}$$

Inverse rotation (r):

$$\begin{cases} x = +x' \cdot cos(r) + y' \cdot sin(r) \\ y = -x' \cdot sin(r) + y' \cdot cos(r) \end{cases} \tag{4.10}$$

Inverse linear tapering along y-axis (t):

$$\begin{cases} x = \frac{x'}{\frac{t \cdot y}{a_y} + 1} \\ y = y' \end{cases} \tag{4.11}$$

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

$$\begin{cases} x = -sign(b) \cdot arctan\left(\frac{x'}{y'-c}\right) \cdot \sqrt{(x')^2 + (y' - \frac{a_y}{b})^2} \\ y = \frac{a_y}{b} - sign(b) \cdot \sqrt{(x')^2 + (y' - \frac{a_y}{b})^2} \end{cases} \tag{4.12}$$

4.3.3.3 Optimization Framework

In the Bayesian framework proposed in Gong et al. [316], the authors assumed that some parameters (those concerning shape) have a Gaussian distribution as prior, $N(\mathbf{p_s})$, whereas others (those concerning pose) have a Uniform distribution as prior, $U(\mathbf{p_p})$. The edge strength likelihood is denoted as E. Then, according to the Bayes rule, the posterior probability can be modeled as in Equation (4.13):

$$Pr(\mathbf{p} \mid E) = \frac{Pr(\mathbf{p}) \cdot Pr(E \mid \mathbf{p})}{Pr(E)} = \frac{Un(\mathbf{p_p}) \cdot N(\mathbf{p_s}) \cdot Pr(E \mid \mathbf{p})}{Pr(E)} \propto Un(\mathbf{p_p}) \cdot N(\mathbf{p_s}) \cdot Pr(E \mid \mathbf{p}) \tag{4.13}$$

This results in optimizing the log-likelihood in Equation (4.14):

$$L = ln(Pr(\mathbf{p_s})) + ln(Pr(E|\mathbf{p})) \tag{4.14}$$

4.3.3.4 Proposed Approach

In the proposed approach, the deformable superellipse is modeled as specified in Section 4.3.3.2-**Deformable Superellipses**. Geometric deformations to the fundamental superellipse shape can be obtained as reported in Section 4.3.3.2-**Geometric Transformations**.

The problem of modeling $Pr(\mathbf{p} \mid E)$, as in Equation (4.13), is that it requires to have prior data on edge maps from images of the same kind of those obtained with the ultrasound device that will be used for carrying out the procedures. When it is not feasible to collect such images in advance, it may be preferable to model $Pr(\mathbf{p} \mid U)$, where U is a set of user-defined points. If the model does not need to make rigid assumptions about U, it can provide a fast and reliable system for achieving prostate gland segmentation with only moderate user interaction and without the need to build a large training set.

Therefore, in the proposed formulation, the posterior probability can be written as reported in Equation (4.15):

$$Pr(\mathbf{p} \mid U) = \frac{Pr(\mathbf{p}) \cdot Pr(U \mid \mathbf{p})}{Pr(U)} \propto Un(\mathbf{p_p}) \cdot N(\mathbf{p_s}) \cdot Pr(U \mid \mathbf{p}) \tag{4.15}$$

The prior about shape parameters can be optimized by maximizing Equation (4.16) [316]:

$$ln(Pr(\mathbf{p_s})) = -\sum_{j}\left[\frac{(p_j - m_j)^2}{2\cdot\sigma_j^2}\right] \tag{4.16}$$

Instead, the likelihood linked to the term $Pr(U \mid \mathbf{p})$ can be maximized by optimizing the energy in Equation (4.17), where U is the set of user-defined points, C is the polygon representing the prostate mask boundaries, d is the point-to-polygon distance, and $E(C;U)$ is the energy function to minimize.

$$E(C;U) = \sum_{(x,y)\in U} d\left(C,(x,y)\right)^2 \tag{4.17}$$

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

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

4.3.3.5 Implementation Details

The general workflow employed for TRUS segmentation is reported in Figure 4.5.

First, the user is asked to select points from at least three slices of the TRUS volume. In every slice, the user has to select a variety of points ranging approximately from 4 to 12, as detailed at the end of Section 4.3.3.4. In order to ease this process for the experiments realized during this research, a JSON interface with the popular 3D Slicer software [344] has been realized for this research work.

The user can enter two types of models when inserting points. The first is the superellipse, whereas the second exploits bidimensional B-splines (as implemented by the method *scipy.interpolate.splprep*). For the purposes of 3D modeling, a superellipse is then fitted to the spline in the second case. For mid-gland slices, the B-spline configuration, especially when 10-12 points are annotated by the user, is the recommended way to proceed. When there are

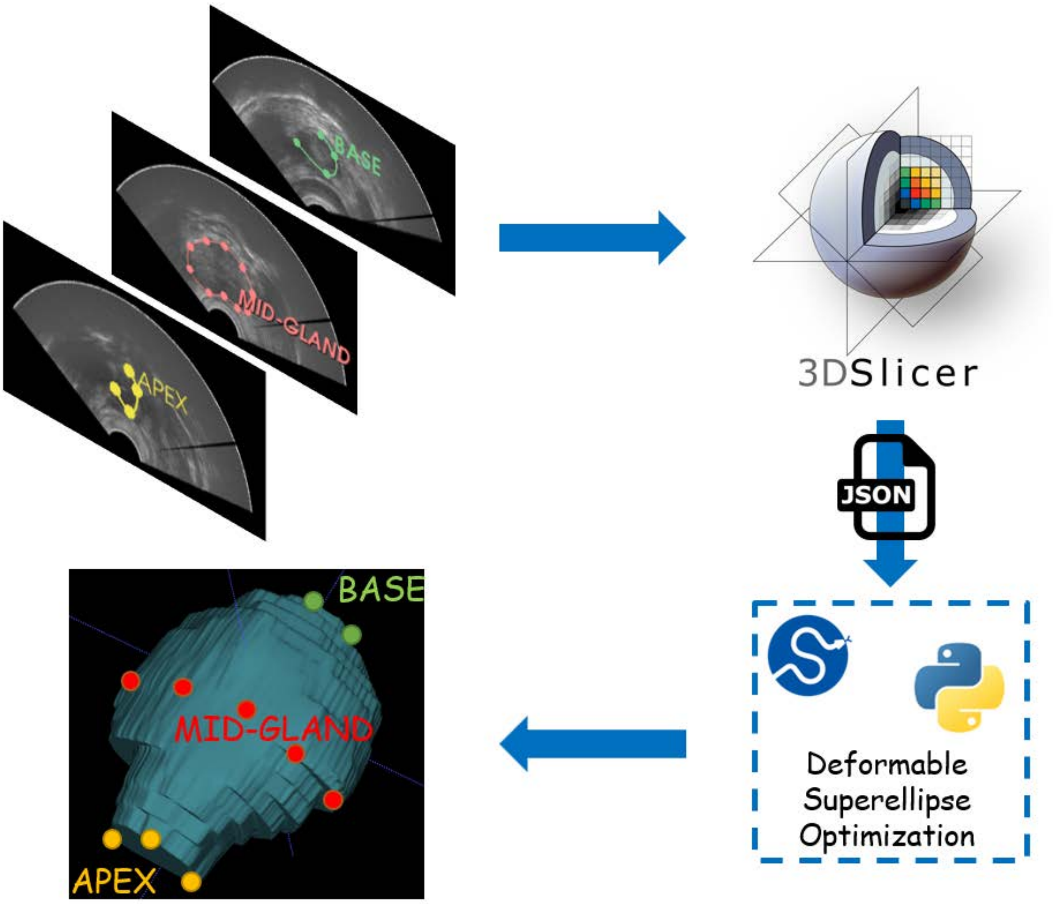


Fig. 4.5 *Workflow for TRUS segmentation with the developed methodology*. The operator needs to annotate some points in at least the apex, base, and mid-gland of the prostate. Then, a JSON file is fed as input to an optimization routine that fits the best 2D superellipse in every annotated slice. Then, a 3D model is built by linearly interpolating 2D models. 3D Slicer has been exploited as GUI to speed up and ease the process.

few annotated points, deformable superellipse is more likely to properly work, considering that it has a relatively low number of parameters. In particular, in the configuration with the least possible number of points, where the user places 4 points at the base, 6 points at mid-gland, and 4 points at the apex, the deformable superellipse should be exploited.

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

After the object is created, the *inside-outside* function reported in Equation (4.3) is used to create a mask of points that satisfies the condition for the centered superellipse. Then, these points are transformed by using deformations in the following order: rotation, as defined in Equation (4.6); linear tapering along y-axis, as defined in Equation (4.7); circular bending along y-axis, as defined in Equation (4.8); translation, as defined in Equation (4.5).

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

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

4.3.4 MRI-TRUS Registration

The described registration algorithm is segmentation-based. Indeed, both MRI and TRUS segmentation masks are required for performing the procedure. Other authors considered this step fundamental too [345, 346]. The particular challenge of MRI-TRUS registration is that the anatomical areas visible in one modality may not be visible in the other.

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

For the binary segmentation mask, the Nearest Neighbour interpolator has been employed to perform the resampling. Output resolution has been set to 0.3mm × 0.3mm × 0.3mm. Segmentation masks have been smoothed with a gaussian kernel with $\sigma = 3$ [346]. Lastly, the Maurer signed distance transformation, which exploits the Euclidean Distance transform

[347], has been applied to the segmentation masks. The steps involved in the registration pre-processing are reported in Figure 4.6.

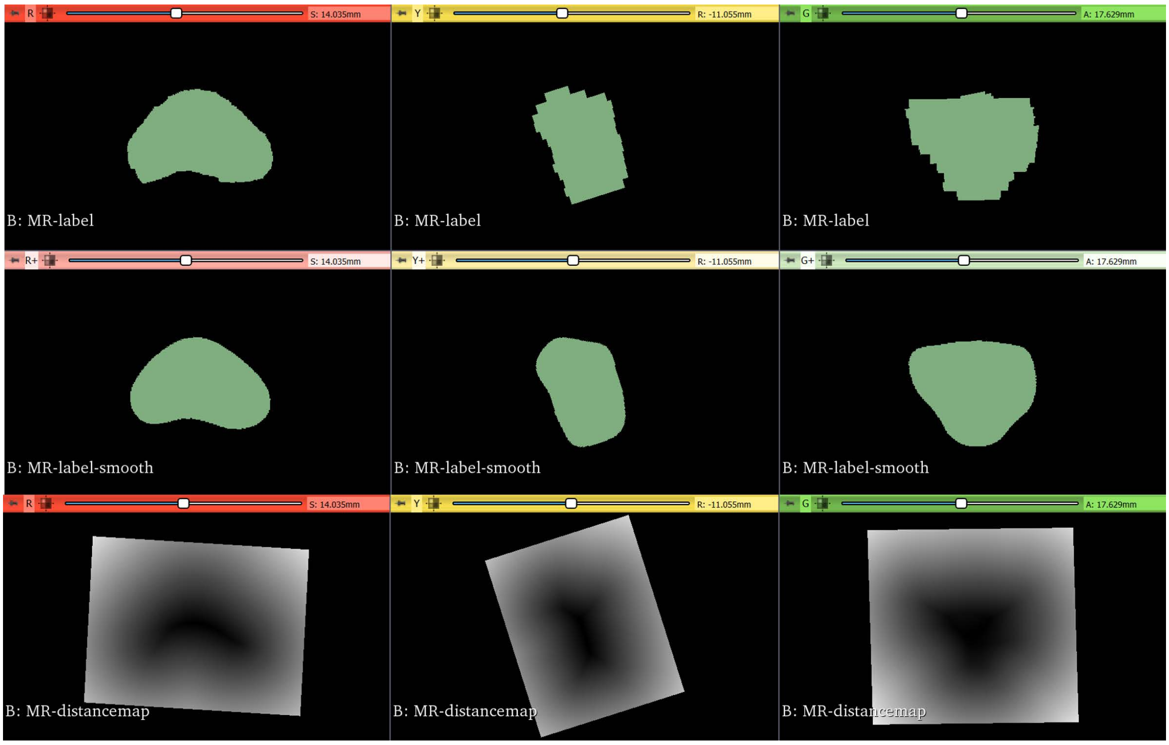


Fig. 4.6 *Image registration pre-processing*. Original ground truth segmentation masks are reported in the top row. Then, they are smoothed with a gaussian filter, as depicted in the middle row. Lastly, Distance Maps are obtained from the smoothed masks, as shown in the bottom row.

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

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

The second approach, instead, determines an initial transformation by considering a set of landmarks. It determines the optimal transform that can map the fixed image and the moving image with respect to the least square errors of the levels of intensity [348].

Since the proposed approach aims to perform the registration of distance maps whose intensity values have the same range of values and meaning, the dissimilarity measure employed is the sum of squares of intensity differences (SSD). Lower values of the said metric correspond to better results. The optimizer employed is based on gradient descent, and is targeted at finding the set of parameters that define a transformation that optimizes the metric as better as possible. The overall workflow employed for the registration with all the various components described in this section is portrayed in Figure 4.7.

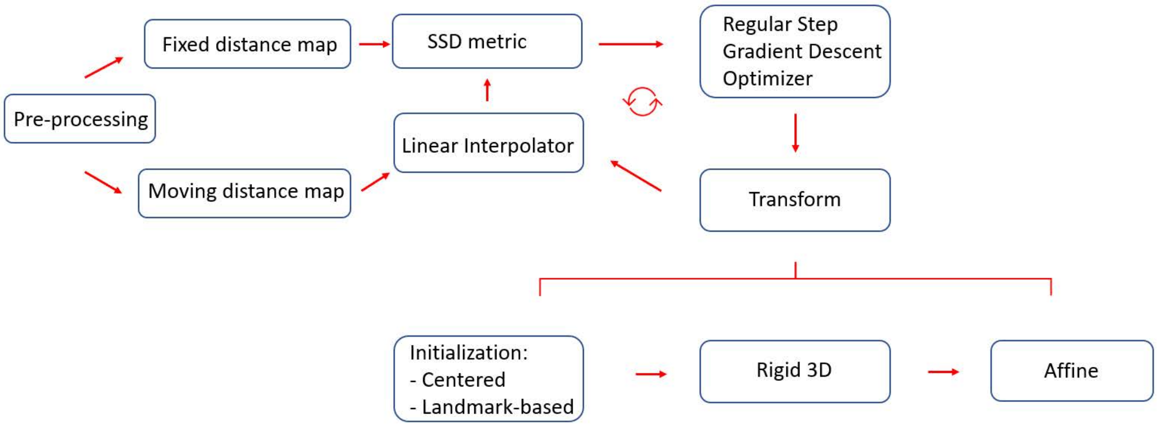


Fig. 4.7 *Proposed registration workflow*. It starts with pre-processing segmentation masks, to make them isotropic and at the same resolution. Thereafter, SSD is exploited as the metric to perform the registration, whereas gradient descent is adopted as the optimizer. Two kinds of initialization have been considered: one based on centers and the other based on landmarks.

4.3.5 Performance Metrics

The performance of the segmentation and registration algorithms analyzed for this study is evaluated by calculating metrics based on the overlap of volumes and metrics based on the distances of the external surfaces points. The metrics used for volumetric overlap require to introduce the predicted volume, P and the ground truth volume, G. They were *Dice Similarity Coefficient* (DSC), *Volume Overlap Error* (VOE), *Relative Volume Difference* (RVD), defined as in Eq. (5.2), Eq. (4.19), and Eq. (4.20).

$$DSC(P,G) = \frac{2 \cdot |P| \cap |G|}{|P| + |G|} \tag{4.18}$$

$$VOE(P,G) = 1 - \frac{|P \cap G|}{|P \cup G|} \tag{4.19}$$

$$RVD(P,G) = \frac{|P| - |G|}{|G|} \tag{4.20}$$

The metrics based on the concept of surface distances include *Hausdorff Distance* (HD) and *Average Symmetric Surface Distance* (ASSD). Definitions for these metrics can be found in [349].

4.4 Results

4.4.1 Segmentation

The results achieved for the segmentation of MRI and TRUS are illustrated below.

4.4.1.1 MRI

Quantitative results for MRI segmentation with nnU-Net are reported in Table 4.2, whereas qualitative results, as segmentation masks, are depicted in Figure 4.8. The nnU-Net model trained on the PROMISE12 challenge has been used for obtaining the best results [297, 313]. The SAML-V dataset has been obtained by sampling 24 images for validation from the SAML dataset. It is worth noting that the Dice coefficient is higher than 88 % and ASSD is less than 1 mm for both validation sets under consideration, showing the reliability of the nnU-Net framework for automatic MRI segmentation of the prostate region.

Table 4.2 *Quantitative metrics results for MRI prostate segmentation with nnU-Net*. Performance has been measured on two validation sets.

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

4.4.1.2 TRUS

Quantitative results for TRUS segmentation with the developed methodology based on deformable superellipses are reported in Table 4.3, whereas sample segmentation images are depicted in Figure 4.9. Three experiments have been conducted for each case, placing 4 points on the base and 4 on the apex, using only the superellipse to fit the contours. Instead, on the mid-gland, a number of points varying from 10 to 12 has been considered, exploiting B-splines before fitting the superellipse to finally achieve the 3D modeling of the prostate gland. Results are reported as mean ± std of the experiments done on each case. It is possible to see that results are overall considerable, being the Dice coefficient greater than 87% in

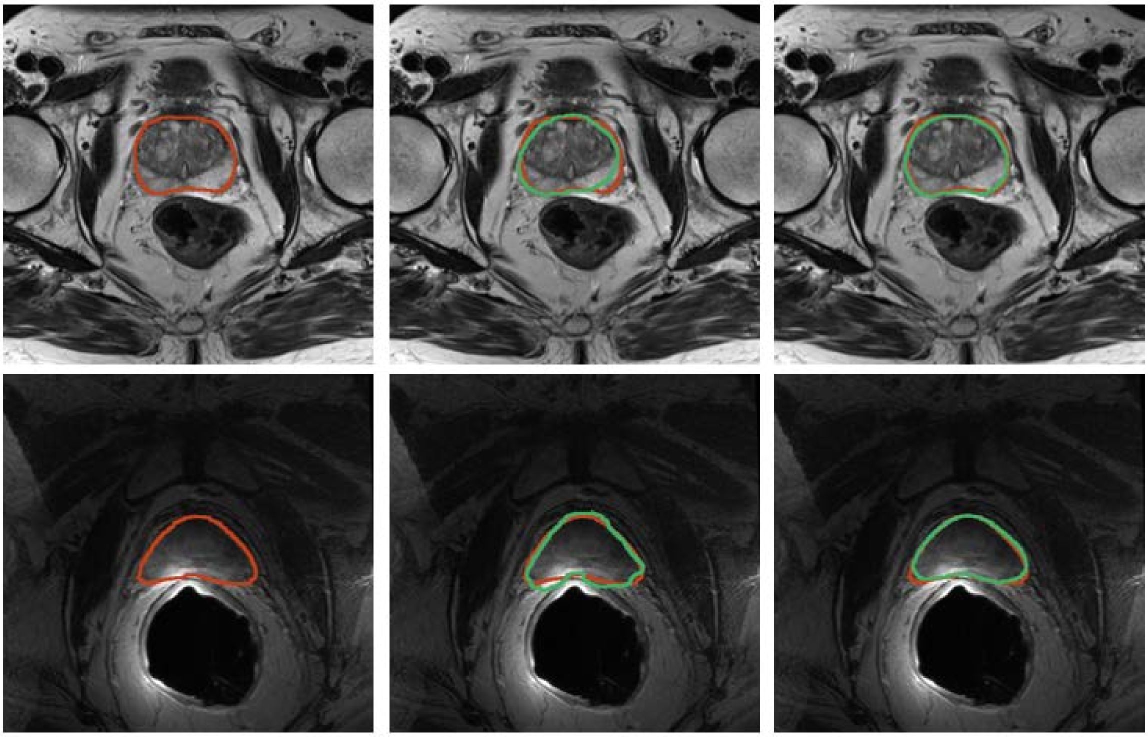


Fig. 4.8 *Results for prostate segmentation from MRI.* The top row contains images from the SAML dataset, whereas the second row encloses slices from ZENODO. The ground truth is represented in red, whereas the predictions from the nnU-Net models are colored in green. The middle image shows the prediction mask for the nnU-Net trained for only 10 epochs, whereas the right image depicts the prediction mask for the one trained on the PROMISE12 dataset.

all cases. Moreover, the proposed implementation is iterative, so that the user can refine the results until it reaches the desired performance. For the purposes of this research, the experiments stopped at the second iteration, which allowed the enhancement of results in all cases.

Table 4.3 *Quantitative metrics results for TRUS prostate segmentation with the proposed superellipse-based approach.* Results are reported for both the 1^{st} and the 2^{nd} iterations of the algorithm execution.

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

4.4.2 Registration

Quantitative results for registration across the two considered imaging modalities, TRUS and MRI, are reported in Table 4.4 for the configurations with and without landmarks, respectively. An example of the workflow for the image fusion is depicted in Figure 4.10. The Dice coefficient is higher than 91 % for all the cases, and HD is less than 4 mm, demonstrating that the developed registration method is promising.

Table 4.4 *Quantitative registration results.* Results are shown in two different configurations. The first exploits as the initializer the center of the images, whereas the second employs a set of landmarks.

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

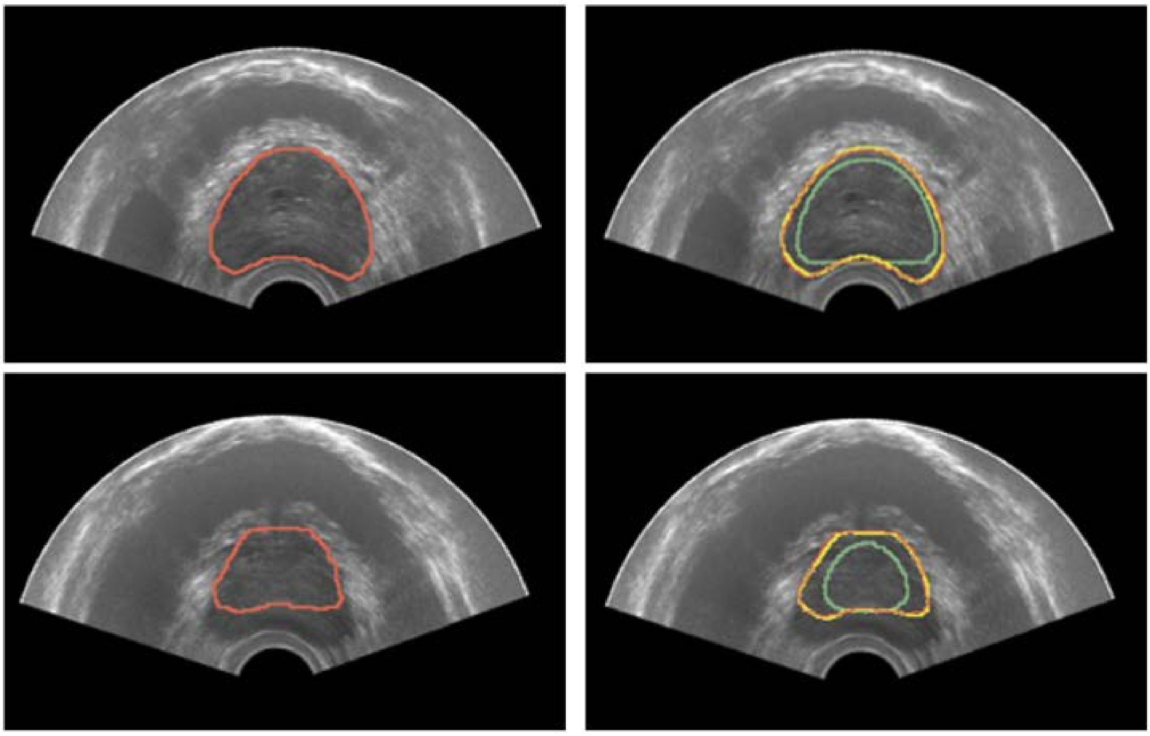


Fig. 4.9 *Qualitative results for prostate segmentation from TRUS*. The left image portrays the ground truth prostate mask in red. The right one instead depicts the segmentation results both after the first and second iteration in green and yellow colors, respectively.

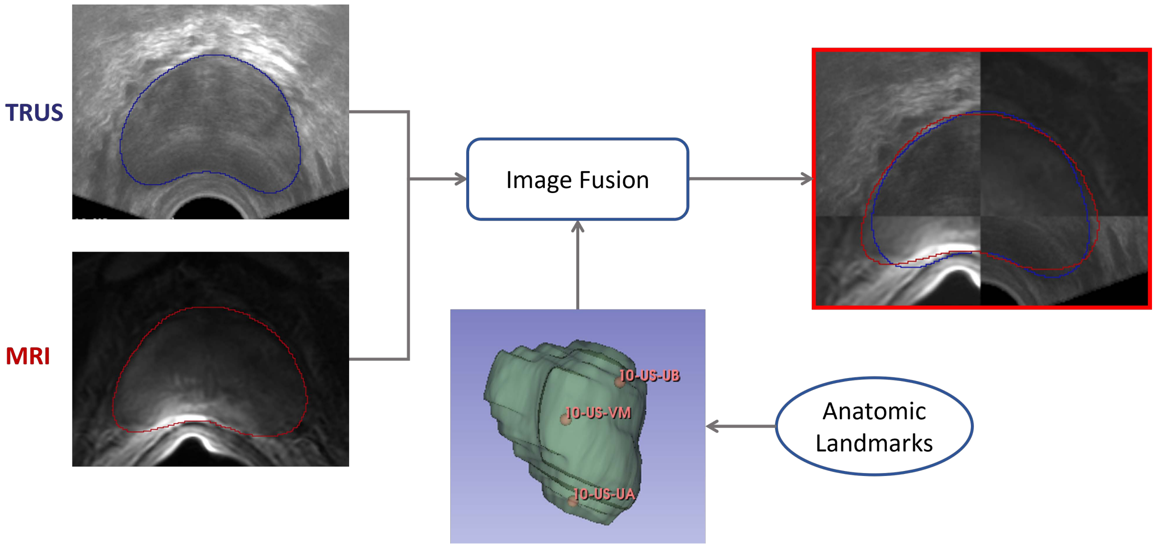


Fig. 4.10 *Image fusion workflow*. Segmentation masks are obtained for both domains: TRUS and MRI. Then, the registration is performed as described in Section 4.3.4, so that images can be fused. Both masks have been shown after the registration procedure.

4.5 Discussion

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

The deformable superellipses are shape models that allow modeling a variety of geometry deformations starting from ellipses, which can resemble the most common prostate shapes. In fact, the prostate shape can be well approximated by a tapered ellipsoid [315]. When the procedure is performed, the transducer induces a slightly posterior deformation in the patient's prostate which can be modeled, for instance with the bending parameter *b*.

Therefore, this work proposed a novel formulation of the deformable superellipse to make it a suitable method for TRUS segmentation, also in the absence of training data from a given transducer. Other approaches, like that of Gong et al. [316], require edge detection algorithms, so that could be exploited for automatic segmentation, but on the other side, demand training data from the specific transducer. The advantage of the proposed method is that it can be applied in any circumstance, only necessitating a moderate interaction with the physician, and always yields considerable results.

In the experiments carried out for this study, the proposed method required 41 ± 7 s for placing the points in three or four slices, whereas it took 5 ± 1 s to build the 3D model. The time needed for the second iteration was more variable—74 ± 32 s. The superellipse implementation of Mahdavi et al. [315] took 32 ± 14 s for initialization, which is similar to the time needed to place the initial user-defined points in the proposed approach. On the computational side, it needed 14 ± 1 s, which is more than the considered implementation. Furthermore, in their case, segmentation refinement can be performed by the user, with a time ranging between 1 and 3 minutes. It is not possible to directly compare the proposed approach with the work of Gong et al. [316] since their method is capable of performing segmentation in less than 5 s per slice, but it only delineates 2D boundaries.

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

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

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

4.6 Summary

Prostate segmentation from MRI and TRUS is a complex challenge but may have a huge impact on clinical setups for fusion biopsy. For what concerns MRI, with the advent of the nnU-Net framework, the challenge is more easily met since a standardized pipeline can be employed for semantic segmentation. However, there is still a lack of substantial data and standardized methodologies for TRUS images. In this thesis section, an approach that can be employed in the absence of training data is proposed; the underlying concept mainly relies on the theory of deformable superellipses. With the only requirement of moderate user interaction, the developed methodology reliably segments the prostate from TRUS images.

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

Chapter 5

Deep Learning for Vertebrae Morphology

5.1 Deep Learning based Vertebrae Identification and Segmentation

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

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

One of the limitations of the proposed methodology is its semi-automaticity. However, it offers potential benefits over simple 3D component analysis of the segmented region, as it may result in the spine being considered a unique connected component. In addition, the unavailability of a dataset large enough is always a considerable problem; thus, the proposed methodology can be useful also when there is little or no annotated data. A tabular comparison of the proposed approach with existing research methods is listed in the Table 5.1. Further comparison information can be found in later sections.

Table 5.1 *Comparison between the proposed method and related works*. Results are reported in terms of Dice coefficient (*DSC*). CT stands for computed tomography, CNN stands for convolutional neural network, PaDBN stands for patch-based deep belief networks, PCNN stands for pulse coupled neural network, APCNN stands for adaptive pulse coupled neural network, MLPNN stands for multi-layer perceptron neural network and MLPNN1f means MLPNN considering only the intensity level of each voxel as a feature.

Reference	**Method**	**Test Sample**	***DSC* [%]**
Proposed	3D V-Net	50 CT scans	89.17 ± 3.63
Kim et al. [350]	U-Net	14 CT scans	90.40
Vania et al. [351]	CNN	32 CT scans	94.28 ± 3.25
Qadri et al. [352]	PaDBN	3 CT scans	86.1
Lessmann et al. [353]	3D U-Net	25 CT scans	84.6 ± 6.9
Zareie et al. [354]	PCNN	17 CT scans	65.7 ± 15.4
	APCNN	17 CT scans	95.0 ± 2.3
	MLPNN	17 CT scans	91.1 ± 2.9
	MLPNN1F	17 CT scans	77.3 ± 4.7
	APCNN (noise 3%)	17 CT scans	94.3 ± 2.6
	MLPNN (noise 3%)	17 CT scans	87.8 ± 4.1

5.2 Introduction and Background

The spine plays a primary role in sustaining and supporting the human body and shielding organ structures while allowing the full body mobility. It also protects the spinal cord from injuries and mechanical shocks due to impacts [355]. The anatomic complexity of the spine, which consists of 33 vertebrae, 23 intervertebral disks, the spinal cord and connecting ribs, often leads to an under-diagnosis of spinal pathologies [356]. The spinal surgeon is faced with the need of robust algorithms to segment and create a spine model, leading to the development of Computer-Assisted Surgery (CAS) systems [351]. The knowledge of the shape of single vertebrae can aid early diagnosis of degenerative disorders, spinal deformities or trauma and support surgical planning [357]. CT is the most spatially accurate imaging modality to assess the three-dimensional morphology of the vertebra [358].

The most significant challenges in the context of vertebrae segmentation and identification, including large-scale vertebrae segmentation challenges (VerSe‘19 and VerSe‘20), have been organized during the MICCAI international conferences [359, 360]. The data of

Verse‘19 are composed of 160 CT scans, whereas that of Verse‘20 comprise 300 CT scans. Previously available datasets from other challenges in the spine imaging domain are much smaller. Examples include the Computational Spine Imaging 2014 Workshop, targeted at the segmentation of the thoraco-lumbar spine, which consists of 20 images [358, 361], and the online challenge xVertSeg, targeted at the segmentation of the lumbar spine, composed of 25 samples [357].

When elaborating spine imaging data, vertebrae classification and vertebrae segmentation are two pivotal tasks. Applications span from diagnosis (detecting and grading of vertebral fractures, spinal curve estimation, identification of spinal deformities), to biomechanical modeling and surgical planning for metal insertions. As a radiological imaging technique, CT scans are the gold standard for assessing the ‘bone’ part of the spine, since they guarantee high bone-to-soft-tissue contrast [355]. Several methods have been proposed in the literature for vertebrae segmentation and labeling. Traditionally, spine segmentation has been approached predominantly as a model-fitting problem; however, more recent spine segmentation techniques focused on DL-based methods [353].

5.3 Related Studies

Kim et al. developed a web-based tool for spine imaging data segmentation from CT scans [350], exploiting deep learning-based methodologies, especially the U-Net architecture [324]. The tool was implemented in Python with the Keras library for the data processing side, whereas a Flask server framework was developed for providing accessibility over the web. The U-Net was trained on 310 images from CT scans, validated on 20 images and tested on only 14 images. This approach allowed the authors to obtain a Dice coefficient of roughly 90% for the binary spine segmentation task.

Vania et al. proposed a method for the automatic spine segmentation from CT scans using CNN via the generation of redundant class labels [351]. The implemented architecture consisted of two convolutional layers and three fully connected ones. Besides classes for background and spine, two redundant classes were generated by dilating the spine mask. This approach allowed the authors to obtain a Dice coefficient of 94% for the binary spine segmentation task.

Qadri et al. developed an automatic approach, named patch-based deep belief networks (PaDBNs), for vertebrae segmentation in CT images [352]. Deep belief networks (DBNs) are DL models composed of stacked Restricted Boltzmann Machines (RBMs) [362]. Their proposed model allowed them to automatically select the features from image patches and measure the difference between classes. Unsupervised learning was exploited for weight initialization, whereas supervised fine-tuning was used to update weights. One strength offered by this methodology is the considerable reduction in the computational cost while retaining good performances.

Zareie et al. [354] proposed two methods for vertebrae segmentation in 3D CT images. The first is based on a multi-layer perceptron neural network (MLPNN), which used seven gray-level statistical features to classify each voxel as vertebrae or background. The second method implemented an Adaptive Pulse Coupled Neural Network (APCNN) to segment vertebrae and used a median filter to refine the results. This network was a modified version of the one proposed by Chang et al. [363], in which the parameters were adjusted adaptively according to the input image. The performances of both systems were calculated in terms of *DSC* on seventeen 3D vertebrae CT images of the thoracic and lumbar spine of both normal and abnormal cases. The results compared four different models: the PCNN developed by Chang et al. [363], the APCNN, the MLPNN with all seven features and the MLPNN using only the intensity level of each voxel as a feature (MLPNN1f). In addition, the robustness of APCNN and MLPNN was evaluated by adding salt-and-pepper noise to the images. It was shown that the APCNN performed better than the other methods with a *DSC* of 95%, being less sensitive to noise than the MLPNN and more adjustable to each image than a classic PCNN.

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

Bae et al. proposed a fully automated approach for 3D segmentation and separation of multiple cervical vertebrae in CT images exploiting a 2D CNN [364]. The authors trained a 2D U-Net model for performing the spine segmentation, considering two classes for the superior and the inferior part of the vertebra, obtaining accuracies comparable with the inter- and intra-observer variability of the manual segmentation performed by human experts. In order to separate the vertebrae, the authors proposed a post-processing stage. In the first part, each class region continuity is assessed by using the connected component analysis to correct mis-labeling errors. Then, a technique for detecting separation points across the superior and inferior region was implemented. After having identified these points, voxels belonging to superior and inferior part of the same vertebrae were merged, leading to a final segmentation with distinct vertebrae.

In the realm of the methods proposed for the VerSe challenges, it is worth considering the top three works from the VerSe‘19 challenge [353, 355, 360, 365] for their accurate results. All the methods that have been proposed for these challenges also considered the vertebrae identification problem.

Sekuboyina et al. [360] proposed a pipeline to localize and identify the vertebrae on multidetector CT scans employing an improved version of the Btrfly Net [366]. The Btrfly fully convolutional neural network works on sagittal and coronal 2D projections and incorporates the spine localization and anatomic prior information using the adversarial discriminators. The approach was tested on both a public dataset of 302 CT scans and two in-house datasets with a total of 238 CT scans. On the public dataset, the network achieved a vertebrae identification rate of 88.5%. On the in-house datasets, instead, with a higher interscan data variability, an identification rate of 85.1% was obtained. One of the principal limitations of the study was that it only considered 24 labels for C1-L5 and did not account for segmentation anomalies, such as L6, or transitional vertebrae, such as a lumbarized S1 vertebra.

Lessman et al. proposed iterative FCNs. The idea was that an instance memory keeps information about previously segmented vertebrae. This memory was then combined with an FCN. This network iteratively analyzed image patches, searched for the first not yet segmented vertebra, which was recognized as completely or partially visible, so that partially visible vertebrae were excluded from further analysis [353].

Payer et al. proposed a cascaded approach which involves three stages. In the first step, a U-Net [324] variation was exploited to regress a heatmap of the spine centerline, which was generated by combining Gaussian heatmaps of all the individual landmarks; this allowed to locate the approximate position of the spine. In the second phase, SpatialConfiguration-Net was employed to localize centers of the vertebrae bodies. It effectively combines local aspect of reference points with their spatial configuration. In the last stage, a U-Net trained with cross-entropy was exploited for the binary segmentation of single separated vertebrae [365]. A more detailed list of approaches considered for the VerSe challenges is presented in [355].

5.4 Materials

The main publicly available datasets concerning spine segmentation are: VerSe‘19 and VerSe‘20 [355, 359, 360], CSI-Label 2014 [367, 368], CSI-Seg 2014 [358, 361] and xVertSeg [357]. The latter three are listed by SpineWeb (http://spineweb.digitalimaginggroup.ca/, *last accessed: 6 June 2021*), an important online archive for multi-modal spine imaging data. A summary of the available datasets is presented in Table 5.2. xVertSeg is a collection of 25 lumbar-only CT scans with voxel-level annotations that include fractured vertebrae. CSI-Seg 2014 and CSI-Label 2014 have become available during the MICCAI 2014. While the former includes segmentation masks, the latter is provided with centroid annotations.

The VerSe‘19 and VerSe‘20 challenges enabled the adoption and benchmarking of deep learning-based techniques for spine segmentation since they have an adequate sample size with a variety of conditions, fields of view, and labeled vertebrae. For instance, the VerSe‘19 data include a variety of fields of view (e.g., cervico-thoraco-lumbar and thoraco-lumbar scans), a compound of isotropic and sagittal reformations, and subjects with metallic implants or vertebral fractures. The Verse‘20 dataset includes atypical anatomies such as transitional vertebrae and other vertebrae, i.e., L6, sacralization of L5 and C7 with cervical ribs. For running the experiments, a dataset of 214 spine multi-detector CT scans has been extracted from VerSe‘20 challenges data. The images can be downloaded from the OSF repository (https://osf.io/t98fz/, *last accessed: 6 June 2021*) and are available in Neuroimaging Informatics Technology Initiative (NIfTI) format. The dataset is split as follows: 148 CT scans for training, 16 CT scans for validation, and 50 CT scans as the final test set. 12 CT scans collected from Medica Sud s.r.l. were also considered, in order to assess the performance of the vertebrae labeling algorithm on patients with scoliosis, ranging from mild to severe cases.

Table 5.2 *Spine segmentation datasets*. In the labels column, C stands for centroids' labels, M for masks, and S for scoliosis severity. Medica Sud s.r.l. (https://www.medicasud.it/, *last accessed: 6 June 2021*) is a local medical clinic that provided 12 CT scans. Please note that this dataset is not publicly available.

Dataset	Spine Tract	Sample Size	Modality	Labels
xVertSeg [357]	Lumbar	$n = 25$	CT scans	C
CSI-Seg 2014 [358, 361]	Thoraco-lumbar	$n = 20$	CT scans	M
CSI-Label 2014 [367, 368]	Whole spine	$n = 302$	CT scans	C
Verse‘19 [355, 359, 360]	Whole spine	$n = 160$	CT scans	C + M
Verse‘20 [355, 359, 360]	Whole spine	$n = 300$	CT scans	C + M
Medica Sud s.r.l.	Whole spine	$n = 12$	CT scans	S

5.5 Methods

DL methodologies are emerging in the medical imaging community for tasks such as segmentation and classification. DL refers to the adoption of computational models with hierarchical level of abstractions capable to jointly extract feature and process them to predict an outcome, and DL methodologies excel in tasks where it is hard to design handcrafted algorithms, as computer vision ones [369]. Convolutional neural networks are a powerful methodology for addressing image segmentation problems, especially with fully convolutional neural networks [370] and encoder-decoder architectures as U-Net [324], U-Net 3D [325], and V-Net [326]. For major details, surveys about U-shaped architectures and semantic segmentation approaches can be considered [322, 371].

To ensure the reproducibility of the algorithms introduced in Sections 5.5.1 and 5.5.2, and the visualization tool presented in Section 5.5.3, the code has been made publicly available on GitHub (https://github.com/Nicolik/Segm_Ident_Vertebrae_CNN_kmeans_knn, *last accessed: 6 June 2021*).

5.5.1 Spine Segmentation

The images have been pre-processed according to the method proposed by Payer et al. [365], which consists of reorienting, smoothing, and clamping. In the work herein described, the clamping has been performed in the range $[-150, 1000]$, instead of $[-1024, 8192]$, since high atomic number structures such as bone have HU values in the range [250, 1000] [372]. Images and the related masks have been cut with the smallest bounding box containing the spine. Images have been resampled to an isotropic resolution of 1 mm, as in the work from Payer et al. [365]. The workflow followed for the spine segmentation is depicted in Figure 5.1.

The binary segmentation stage is performed by exploiting the V-Net architecture proposed by Milletari et al. [326]. The Dice loss function formulation adopted is the same adopted in the work of Altini et al. [327]. V-Net is an encoder–decoder architecture. The first part of the network is devoted to the feature extraction process; the second reconstructs high resolutions masks, exploiting also skip connections across the encoder–decoder paths. Compared to U-Net, it is worth noting that V-Net exploits down-convolutions, with stride $2 \times 2 \times 2$ and kernel-size $2 \times 2 \times 2$, instead of fixed downsampling realized by $2 \times 2 \times 2$ max-pooling, adding trainable parameters also in this stage. As differences between the original V-Net architecture and that implemented for this thesis, there is the adoption of ReLU non-linearities instead of PReLU ones [373] and the adding of a Batch-Normalization

(BN) layer after each convolutional layer, as also done in Altini et al. [327] and Shen et al. [374].

The network has been trained for 500 epochs with data augmentation (random noise with $\mu = 0, \sigma^2 \in [0, 0.1], p = 0.1$). Every 100 epochs, the trained model has been saved to check for overfitting issues. With the aim to implement the patch-based pipeline and the augmentation operation, the TorchIO library has been utilized [375]. The parameters for the patch sampler were: uniform sampling, 8 patches per volume, and patch size of $64 \times 64 \times 64$. The original V-Net architecture processed input volumes of $128 \times 128 \times 64$, but it was decided to limit the dimension for allowing larger batches to benefit from BN layers.

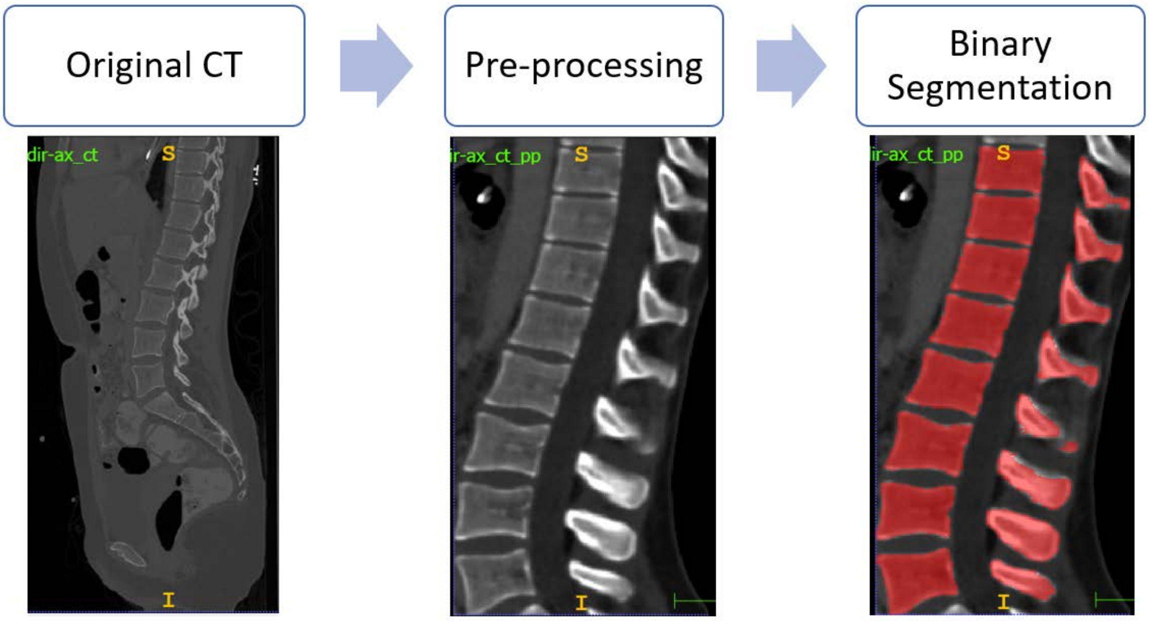


Fig. 5.1 *Vertebrae segmentation workflow.*

5.5.2 Vertebrae Identification

Determining vertebrae centroids is a prerequisite for achieving vertebrae identification. Indeed, centroids can be then exploited to recover single vertebrae masks. This task has to be carried out after the binary segmentation of the spine. A 3D connected component analysis of the segmented region will not solve the problem of vertebrae labeling, since the spine will likely form a unique connected component. For this reason, the chosen approach is a semi-automated workflow that consists of different steps, two of which require input from the user. The workflow followed for vertebrae identification is depicted in Figure 5.2.

The sequence of the involved operations is listed below:

- **Vertebrae Number Selection**. This step requires input by the user, which has to insert the number of visible vertebrae given the binary segmentation. The user has to provide the name of the first vertebra (from the top to the bottom) in order to perform the correct labeling according to the legends provided by VerSe.
- **Slice Extraction**. The algorithm extracts a 2D sagittal slice from the binary segmentation, starting from the middle of the image since it has a higher probability of showing well-clustered vertebrae. Kindly note that, in some cases, e.g., patients affected by severe scoliosis, this consideration may not hold, resulting in lower segmentation performances. Inside the selected slice, the following sub-steps have been carried out:
  - **Morphological and Connected Components Analysis**. The morphological analysis aims at removing small points which can be wrongly considered as standalone components, whereas the purpose of the connected components' analysis is to label each component with a different value.
  - **Shape Descriptor and Clustering for Arches and Bodies**. It is worth noting that every single component is either a vertebral body or a vertebral arch, so it is important to correctly assign each component to the appropriate category. This stage carries out the above process by considering the proper shape descriptors of the individual components.
  - **Arch/Body Coupling**. This step connects each vertebral arch to the nearest vertebral body, by assigning the same label to both.
  - **Centroids' Computation and Slice Showing**. If the output vertebrae number matches the input number from the first step, the algorithm goes further with the computation of centroids' positions for each vertebra; otherwise, the process has to be repeated from another slice.
- **Best Slice Selection and Centroids' Storage**. The algorithm repeats the workflow until it reaches a slice without connected components. Then, the user chooses the best slice among the displayed ones, and the algorithm stores the centroids' position.
- **3D Multi-class Segmentation**. Centroids are used in a *k*-NN classifier to produce a 3D segmentation map in which each vertebra has its own label.

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

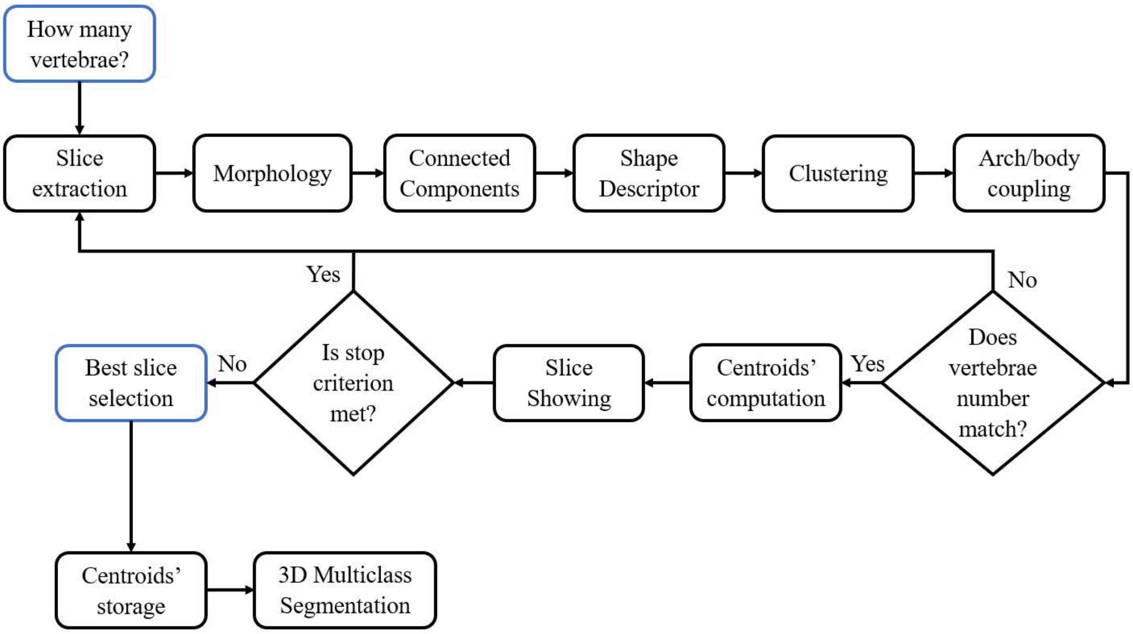


Fig. 5.2 *Vertebrae identification workflow*. Steps that require user interaction have blue borders.

5.5.2.1 Morphological and Connected Components Analysis

As mentioned in Section 5.5.2, the morphological analysis is targeted at removing outliers, i.e., small points that are not connected to vertebral arches or bodies. To carry out this process, the *ErodeObjectMorphologyImageFilter* and the *DilateObjectMorphologyImageFilter* were exploited, which are provided by the SimpleITK library.

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

Figure 5.3 shows the effects of the morphological analysis.

The connected components analysis substage employs the SimpleITK library too, which offers the *ConnectedComponentImageFilter* that labels each component with a different intensity value. Figure 5.4 shows the output of this filter.

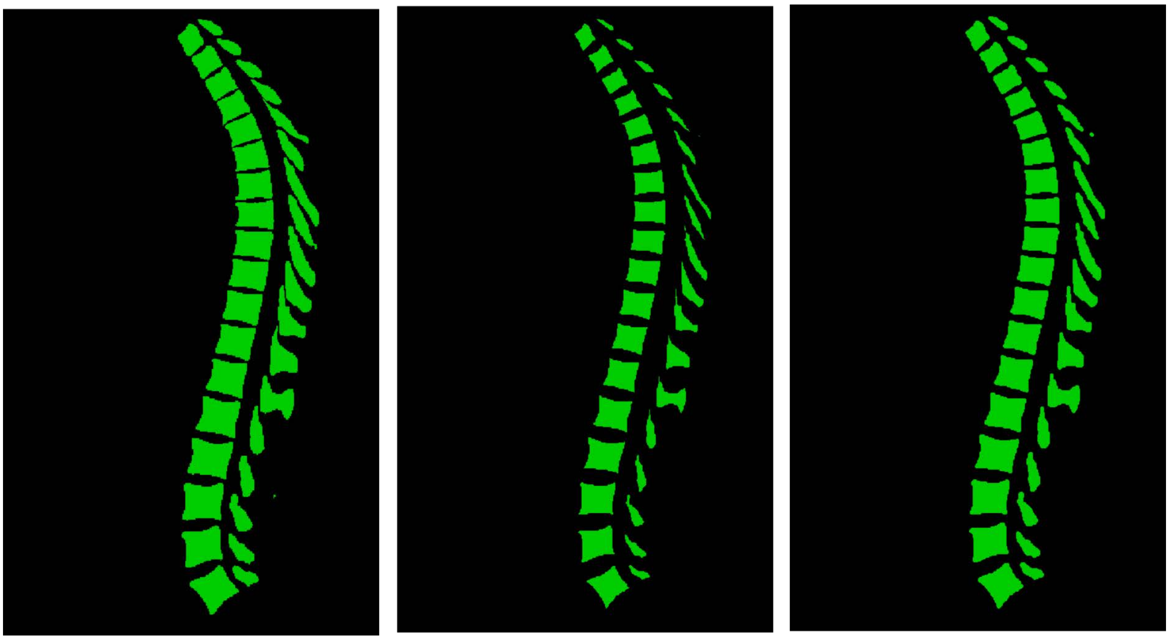


Fig. 5.3 *Workflow of the morphological analysis*. Note that small components are removed in the erosion phase, and they do not appear in the final output.

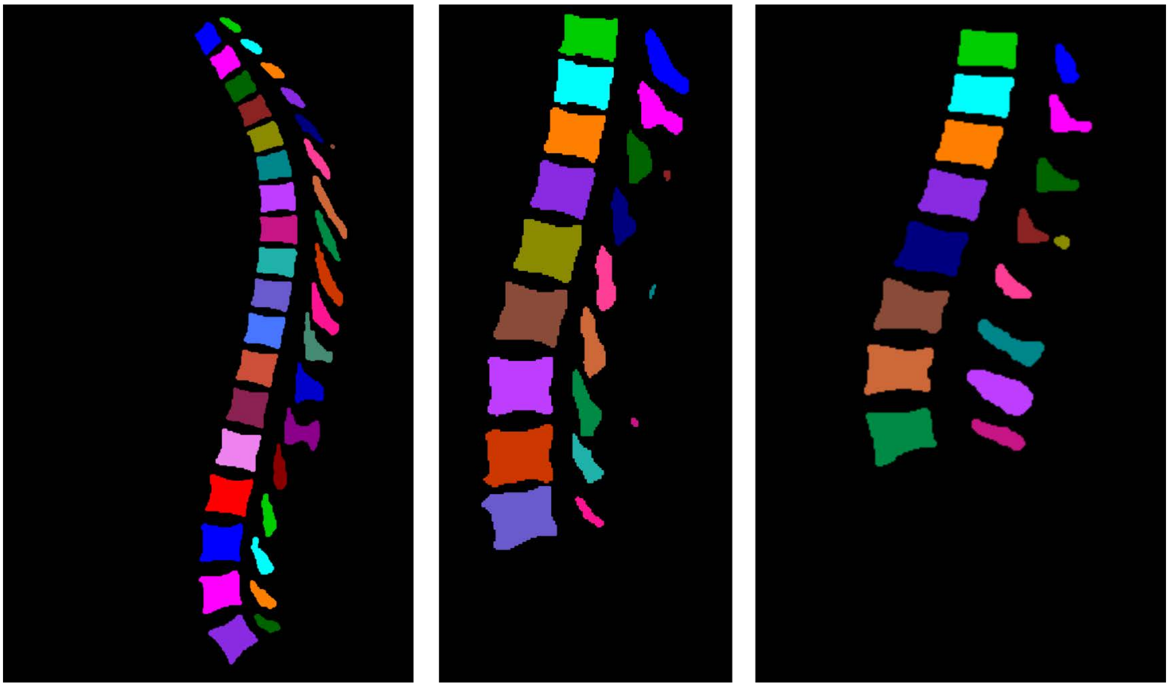


Fig. 5.4 *Sample ConnectedComponentImageFilter output*.

5.5.2.2 Shape Descriptor and Clustering

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

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

- **Area**: *The number of pixels of the region.*
- **Centroid**: *The centroid's position of the region.*
- **Extent**: *Ratio of pixels in the region to pixels in the total bounding box of the region.*
- **Perimeter**: *Approximation of the perimeter of the region.*
- **Eccentricity**: *Ratio of the focal distance (distance between focal points) over the major axis length.*
- **Solidity**: *Ratio of pixels in the region to pixels of the convex hull image (the smallest convex polygon that encloses the region).*

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

The k-means clustering algorithm was exploited for realizing the clustering of vertebral arches and bodies. The drawback is that it is sensible to outliers; however, after the morphological analysis, the probability of retaining outliers is largely reduced. The k-means clustering is a popular algorithm for cluster analysis that finds commonalities in data and groups them without the need for ground truth. Cluster analysis, in fact, belongs to the unsupervised learning branch of machine learning. At the start of the algorithm, k centroids are initialized (randomly, from k data points or from other prior knowledge) in the feature hyperspace. Then, every sample is assigned to the nearest centroid according to some distance metric. Centroids for the next iteration are computed as the average of the coordinates of the points of each cluster. The process is repeated until convergence [376].

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

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

5.5.2.3 Arch/Body Coupling

As mentioned before, this step assigns each vertebral arch to the nearest vertebral body. For each vertebral arch, the Euclidean distance between the arch and every vertebral body is computed. Let N_v be the number of vertebrae and N_a the number of segmented arches. Then, the sets of vertebral arches and bodies can be denoted as $\{a_j\}_{N_a}$ and $\{v_i\}_{N_v}$, respectively, with $j = 1,...,N_a$ and $i = 1,...,N_v$. Distance between pairs of a_j, v_i can be defined as in Equation (5.1).

$$d(a_j, v_i) = \sqrt{(a_{jx} - v_{ix})^2 + (a_{jy} - v_{iy})^2 + (a_{jz} - v_{iz})^2} \tag{5.1}$$

First, the distance between every a_j and v_i is computed. Then, iteratively, the vertebral body which has the minimum distance from the arch is linked to the vertebra body, and the labels are merged. Figure 5.5 depicts the output of cluster analysis and coupling.

5.5.2.4 Multi-class Segmentation

Once the centroids have been computed, every point of the volume which is not part of the background is assigned to the same label as the nearest centroid according to Euclidean

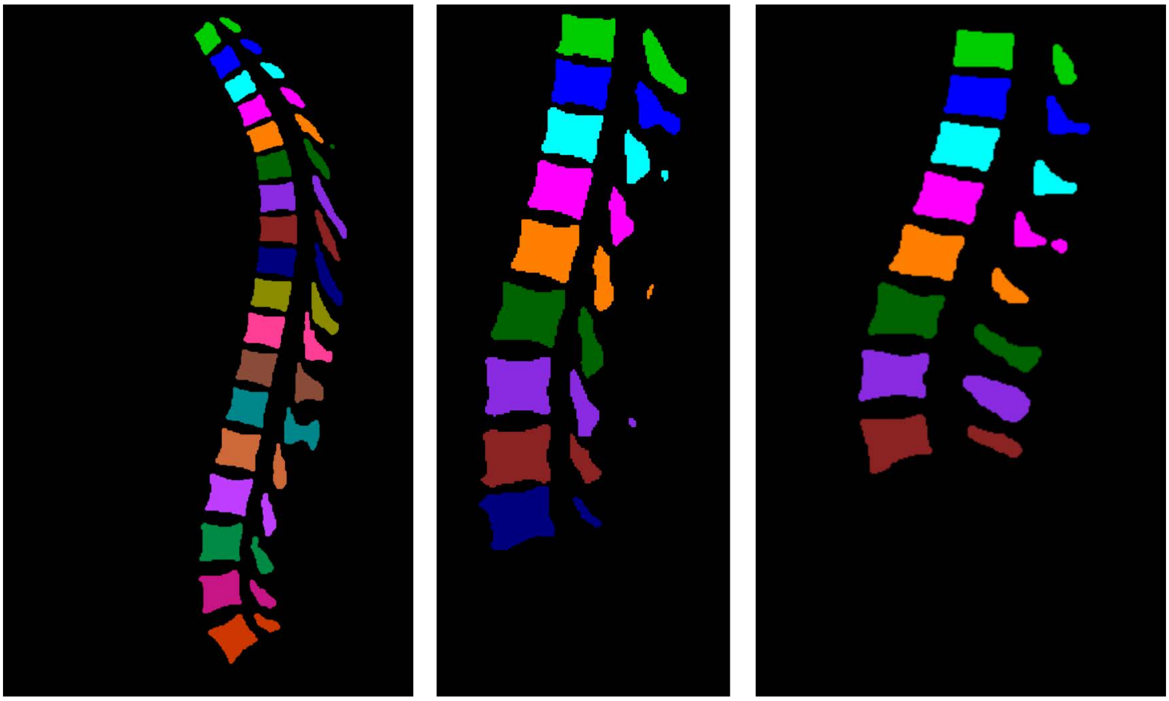


Fig. 5.5 *Sample outputs of cluster analysis and arch/body coupling.*

distance. To accomplish this procedure, a non-parametric method, the *k*-NN classifier, which is based on distances between samples, was exploited. The workflow of this algorithm can be summarized into three main steps [377]:

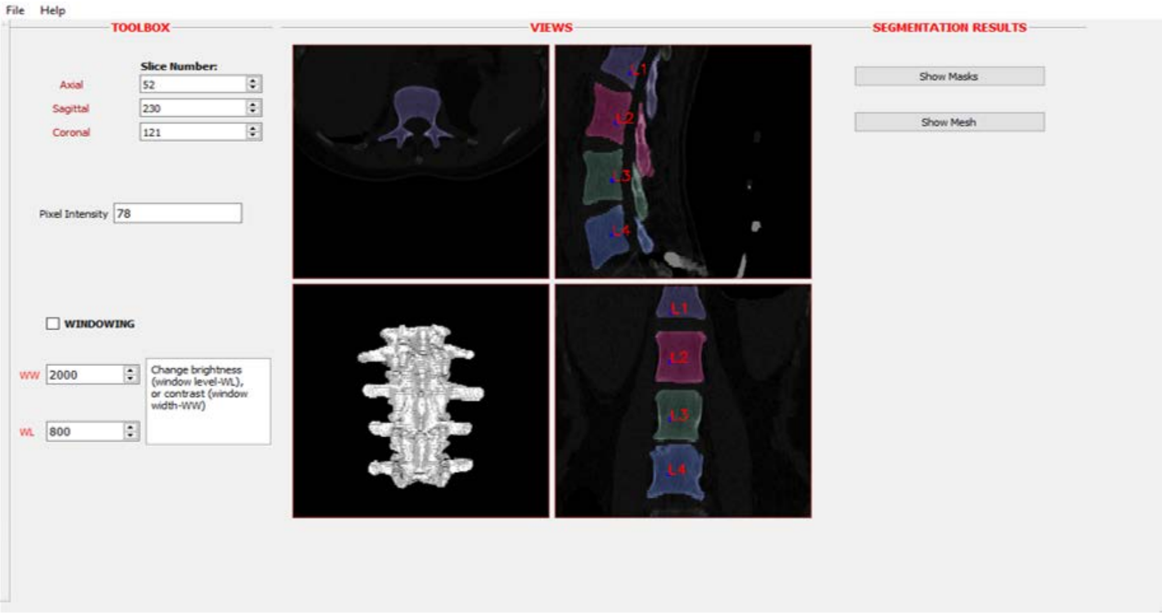
- The learning phase, which is not mandatory, results in the partitioning of the hyperspace in clusters based on samples' positions.
- The distance computation phase consists of the computation of all the distances between samples and centroids (the most used distance metric is the Euclidean Distance,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

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

5.5.3 Visualization Tool

In this study, a visualization tool, that enables displaying the CT scans and masks through a graphic user interface (GUI) based on the Qt framework, ITK, VTK, and OpenCV libraries [378–380], was also developed. It offers multiple functionalities, such as image contrast adjustment, mask visualization, and mesh reconstruction.

As depicted in Figure 5.6, the interface is composed of three major parts. The toolbox section facilitates the loading and reading of the CT images from a local folder and permits smoothly navigating through the slices. The toolbox comprises a windowing option to set custom values of window-width and window-level and adjusts image contrast for optimized visualization of the anatomical regions of interest. The segmentation results are overlaid on the original images and each vertebra is colored differently so that identifying vertebrae is convenient for the user. Furthermore, the vertebrae's centroids and names are shown in the sagittal and coronal views. Three of the four windows in the views section show the CT scan in the axial, sagittal, and coronal planes. A fourth window is used to show a reconstruction of the volume obtained from the segmentation masks. The aforementioned task is achieved by using the VTK discrete marching cubes algorithm.



Slice Number:
Axial 52
Sagittal 230
Coronal 121

Pixel Intensity 78

☐ WINDOWING

WW 2000
WL 800

Change brightness (window level-WL), or contrast (window width-WW)

Fig. 5.6 *Visualization tool for visualizing CT scans and masks.*

5.6 Results

5.6.1 Quality Measures

The quality measures considered, analogously to those adopted in other segmentation works [381–383], can be grouped in two classes:

- measures based on volumetric overlap, such as Dice Coefficient (*DSC*), *Precision* and *Recall*. Such metrics help to compute a similarity degree between the prediction and the ground truth;
- measures based on the concept of surface distance, such as Maximum Symmetric Surface Distance (*MSSD*) and Average Symmetric Surface Distance (*ASSD*).

DSC, *Precision* and *Recall* can be expressed in terms of True Positives (TP), False Negatives (FN) and False Positives (FP). They are defined in Eq. (5.2), Eq. (5.3) and Eq. (5.4), respectively.

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{5.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

In surgical planning applications, it is important to have precise meshes of the anatomical site of interest. In order to properly assess this aspect, a collection of really important quality measures are the one based on the concept of external surface distance. Let P be the predicted volume and G the ground truth volume, then *ASSD* can be defined as in Eq. (5.5).

$$ASSD(P,G) = \frac{1}{|S(P) + S(G)|}\left(\sum_{s_P \in S(P)} d(s_P, S(G)) + \sum_{s_G \in S(G)} d(s_G, S(P))\right) \tag{5.5}$$

where d is a distance measure and $d(s_P, S(G))$ ($d(s_G, S(P))$) is the distance between every point on the surface of the prediction and the ground truth surface (every point on the surface of the ground truth and the prediction surface). The distance between a point and a surface is defined as in Eq. (5.6).

$$d(s_P, S(G)) = \min_{s_G \in S(G)} ||s_P - s_G|| \tag{5.6}$$

The MSSD is defined as in Eq. (5.7).

$$MSSD(P,G) = \max\{h(S(P),S(G)),h(S(G),S(P))\} \tag{5.7}$$

In Eq. (5.7), $h(S(P),S(G))$ denotes the one side Hausdorff distance between the surface of the predicted volume $S(P)$ and the surface of the ground truth volume $S(G)$, as defined in Eq. (5.8):

$$h(S(P),S(G)) = \sup_{s_P \in S(P)} \{ \inf_{s_G \in S(G)} d(s_P, s_G)\} \tag{5.8}$$

5.6.2 Experimental Results

Regarding the binary segmentation task, 3D V-Net obtained the results reported in Table 5.3 considering a test set of 50 images. Quality measures are reported for the CNN trained every 100 epochs, showing that convergence for the considered model can be achieved after 200 epochs. These results allowed the achievement of a realistic segmentation of the whole spine, as can be seen from Figure 5.7.

For what concerns the multi-class vertebrae identification, a multi-class *DSC* of 90.09 ± 3.14% was obtained. The multi-class *DSC* score for the vertebrae labeling task was calculated in an ideal condition, i.e., starting the vertebrae identification procedure from ground truths of spine segmentation. Results are reported as mean ± standard deviation. Multi-class *DSC* is computed as the average of the binary *DSC*s for each vertebra class. Kindly note that the vertebrae identification algorithm proposed in this section may not work optimally for the C1, C2, and C3 vertebrae, since the shapes of the bodies of these vertebrae can be easily mistaken with their arches. The centroids' predictions examples are shown in Figure 5.8. The following vertebrae labeling, obtained with k-NN, is depicted in Figure 5.9.

In order to assess the quality of the vertebrae labeling stage on scoliosis cases, a collection of 12 CT scans coming from Medica Sud s.r.l. was exploited. Four out of the 12 CT scans belong to patients with severe scoliosis, 4 to patients with moderate scoliosis, and the remaining 4 to patients with mild scoliosis. For these images only information about the severity of the scoliosis was available. Therefore, the method of Payer et al. [365] was exploited as the gold standard for the segmentation masks. The multi-class *DSC* has been considered as a similarity measure across the two methodologies. The same binary spine segmentation was achieved for both algorithms, obtained by exploiting Payer's pre-trained models, therefore focusing the comparison on the labeling stage.

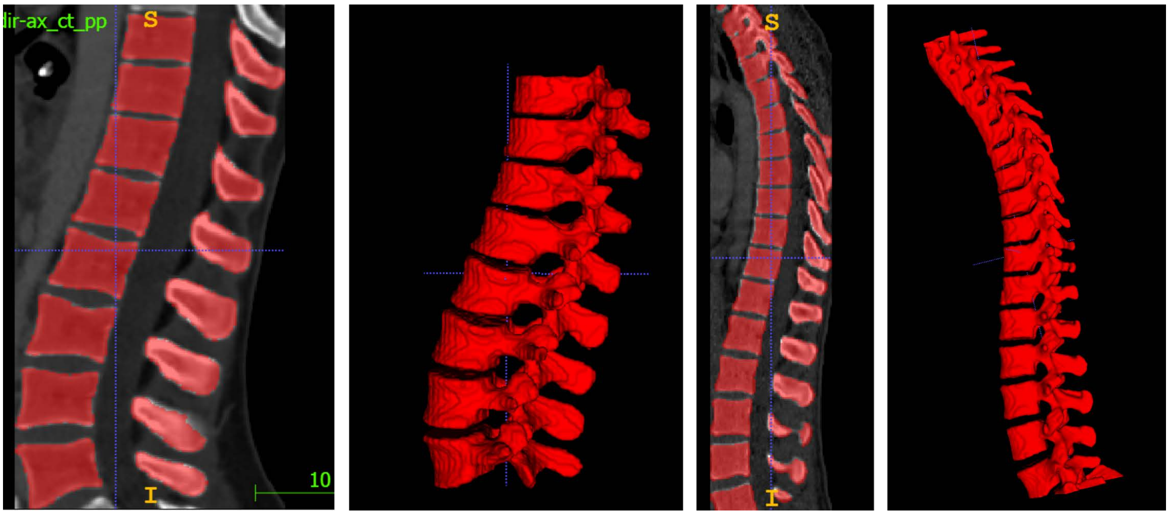


Fig. 5.7 *Examples of binary spine segmentation.*

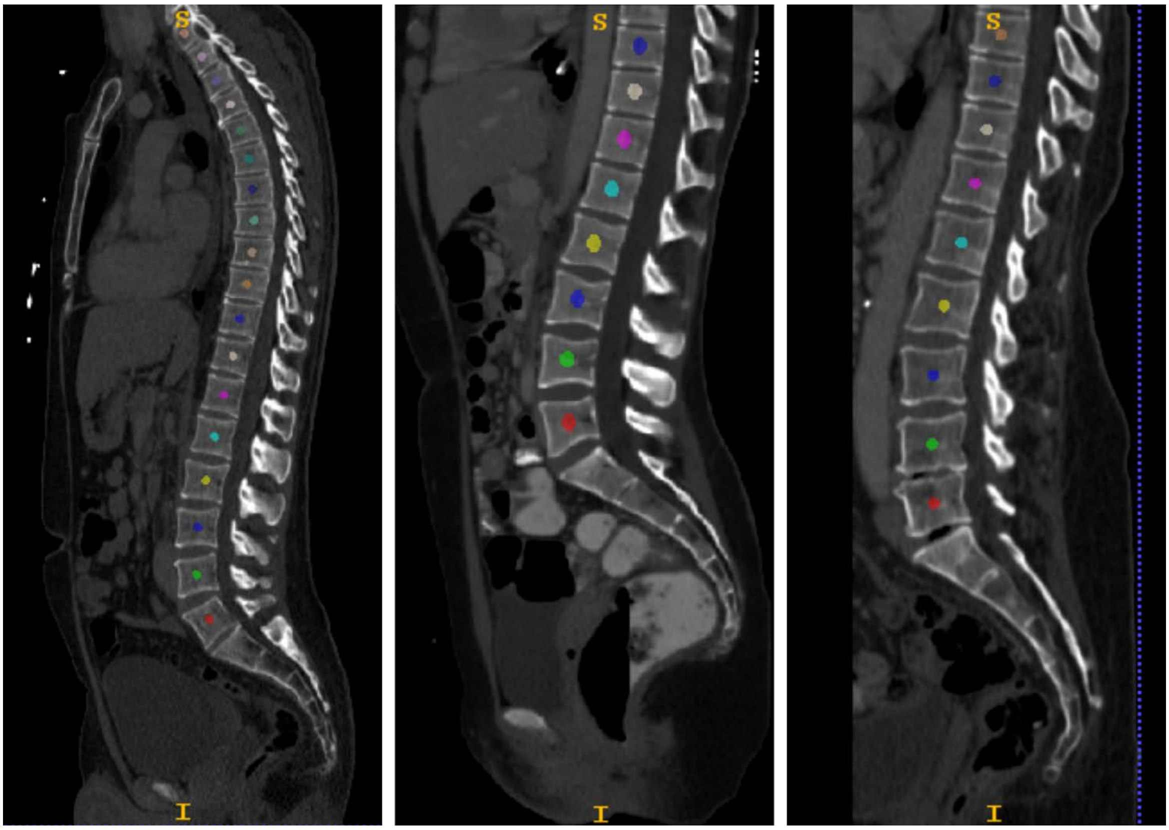


Fig. 5.8 *Samples of the final output of the centroids' prediction.*

Table 5.3 *Spine binary segmentation results*. *ASSD* stands for average symmetric surface distance, whereas *MSSD* stands for maximum symmetric surface distance. The best result for each of the reported measures is in bold font.

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

Multi-class *DSC*s have been determined separately for each scoliosis category, obtaining 43.40 ± 30.03%, 70.61 ± 18.50%, and 83.38 ± 12.51%, for severe, moderate, and mild scoliosis patients, respectively. After adding **Inertia Tensor** to the feature set for discriminating between arches and bodies, multi-class *DSC*s grown, managing to obtain 49.79 ± 24.90%, 77.76 ± 15.05%, and 83.48 ± 12.56%. These results show a clear improvement for severe and moderate scoliosis cases.

The proposed vertebrae identification stage appears to have promising results on mild and moderate scoliosis; however, the identification phase for severe scoliosis cases can be further improved in future works. This problem arises from the fact that the *k*-NN model has a too low complexity for modeling severe scoliosis cases. Example results on this dataset are shown in Figure 5.10. A comparison between the proposed method and related works is reported in Table 5.1.

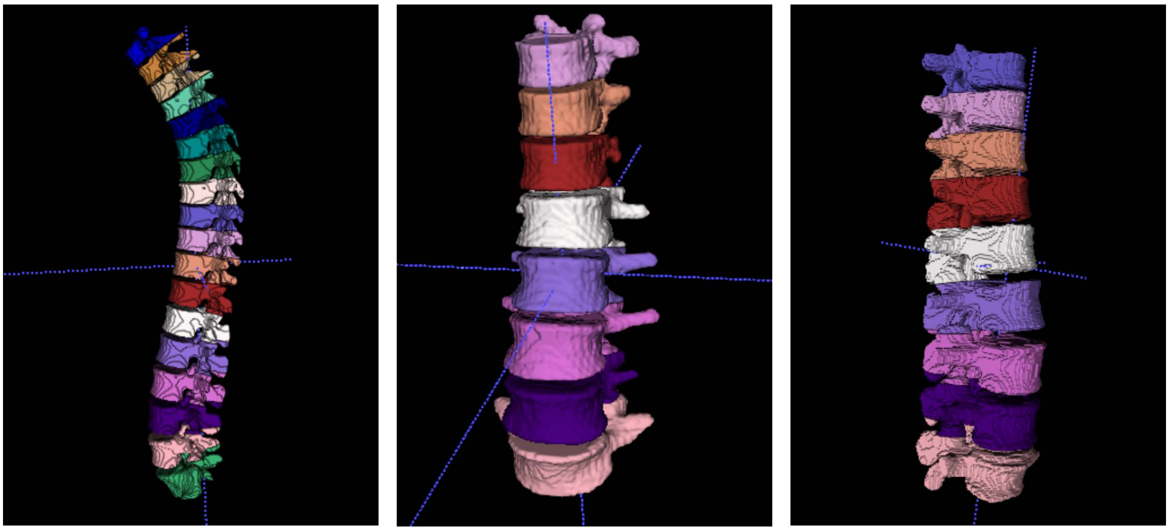


Fig. 5.9 *Sample outputs of meshes of multi-class vertebrae segmentation*. These meshes have been obtained exploiting ITK-Snap (http://www.itksnap.org/, *last accessed: 6 June 2021*).

5.7 Summary

Several methods have been proposed in the literature for the purpose of vertebrae segmentation and labeling. Many of these address only the binary spine segmentation, as in Kim et al. [350] and Vania et al. [351]. However, the methods which involve individual vertebrae segmentation are characterized by complex architectures and require tons of labeled data to be correctly implemented, as for the top scoring architectures in VerSe challenges, e.g., Payer et al. [365] and Lessman et al. [353].

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

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

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

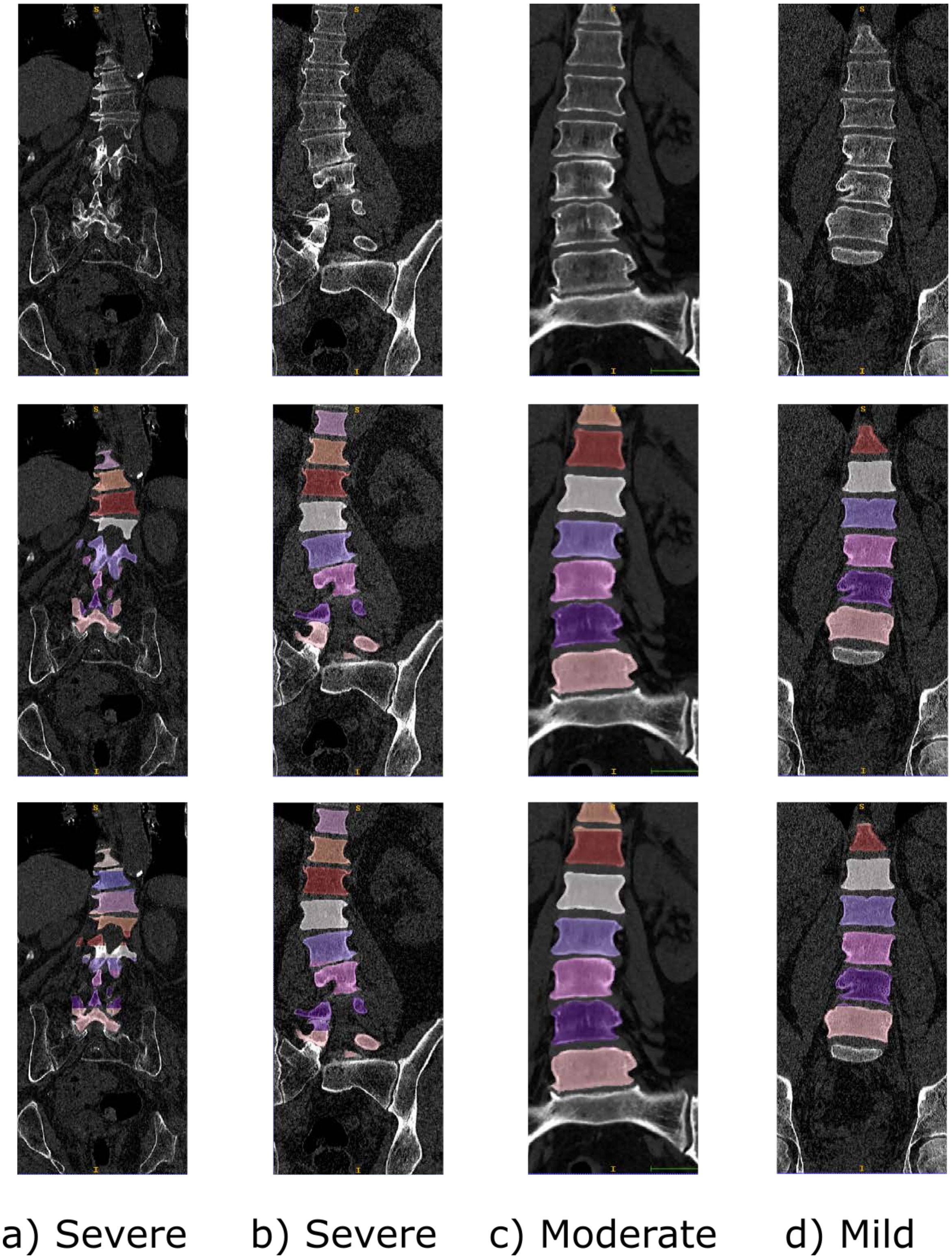


Fig. 5.10 *Comparison between Payer et al. and the proposed method for the vertebrae labeling task*. The first row shows original CT scans, the second row shows predictions from Payer et al. [365], and the third row shows predictions with the proposed vertebrae labeling method. Columns: (**a**,**b**) are of patients with severe scoliosis, (**c**) of a patient with moderate scoliosis and (**d**) of a patient with mild scoliosis. It is important to note that for case (**a**) the proposed method did not provide an accurate result.

Part II

Computer-aided Diagnosis and Explainable Deep Learning

Chapter 6

Deep Learning based Breast Cancer Morphology and Classification

6.1 Breast Shape Classification using CNN

In this study, a novel CNN based DL framework for the classification of breast lesions according to the shape by analyzing the related RoI on DBT images is designed and validated. Considering the shapes of cancerous masses, the Breast Imaging Reporting And Data System (BIRADS) classification of the American College of Radiology, which is the most commonly employed methodology in the clinical and digital breast tomosynthesis settings, has been considered [43]. Such kind of taxonomy refers to the following three classes (see Figure 6.1):

- Regular opacity (Oro) which includes the round, oval, and lobulated shapes;
- Irregular opacity (Ori);
- Architectural distortion shape (Ost).

The clinical importance of the three BIRADS classes consists of the possibility of identifying regular masses or irregular masses/architectural distortions which is the principal purpose of the clinical breast setting for early diagnosis of breast cancer. In fact, it is well known that the *Oro* lesions are usually benign, whereas *Ori* and *Ost* lesions are malignant. Finally, it is also worth mentioning that in this study the *None* class, i.e., images that do not contain any lesion, is also included (see Figure 6.1).

Moreover, the study employs eight state-of-the-art pretrained CNN architectures that have been compared both with and without fine-tuning. Two different online data augmentation routines have been tested to study the impact of several augmentation methods on the

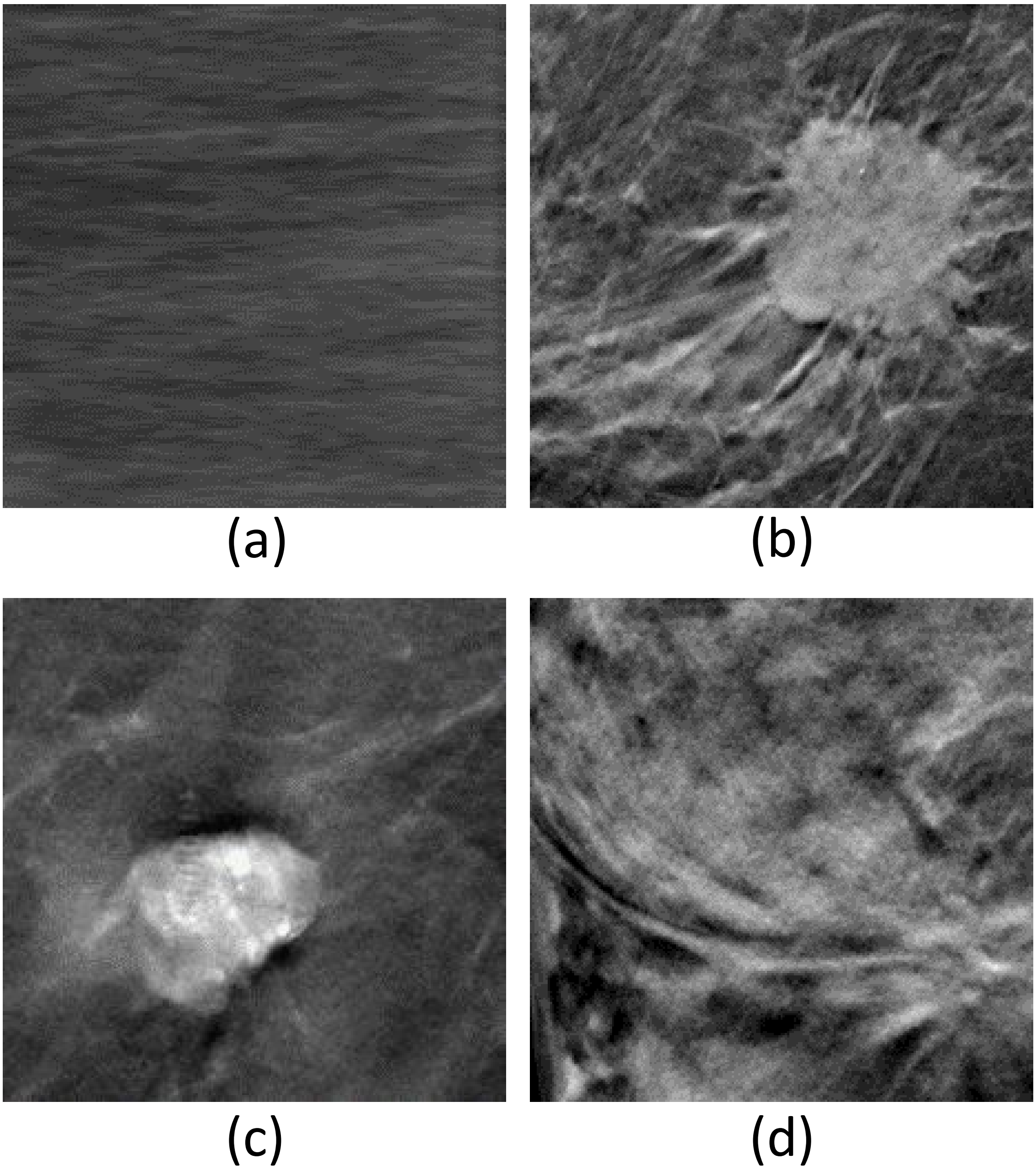


Fig. 6.1 The ready-to-classify RoIs on the images. (**a**) Example of image with no lesions (None); (**b**) example of image with irregular opacity (Ori); (**c**) example of image with regular opacity (Oro); and (**d**) example of image with stellar opacity (Ost).

performances. The dataset used in this study comprises 39 breast DBT exams of 16 patients. Interested readers are kindly referred to such study to explore more about the data acquisition and composition [384].

6.2 Introduction and Background

Breast cancer, which is the second most widespread cancer among women worldwide, has turned into global public health concern due to its complex intrinsic aetiology [385]. The early diagnosis and monitoring of the cancer significantly reduces the death risks, leads to better prognosis and therapy, and lowers the treatment cost.

Mammography wears the crown of being the gold standard among several imaging modalities, because it offers the potential of early detection of pathology [386]. On the other hand, mammography is a 2D method that reduces the ability to visualize lesions in case of prevalent glandular component in dense breast. Moreover, the mammography represents a 2D projection of a 3D structure for which, geometrically, tissues belonging to different planes are superimposed in the radiographic image.

Other imaging techniques including MRI, CT, and DBT are strong candidates where in-depth analysis of hazardous cases is required. Among these, the DBT is proven to have higher accuracy with respect to the 2D imaging methods [9]. After acquisition of the multiple thin and high-resolution images, the DBT system produces a quasi-three-dimensional format of the reconstructed breast images aiming to reduce the effect of tissue superimposition.

Additionally, the required radiation dose is not high, contrary to the conventional imaging techniques, and the generated images appear to have greater resolution and contrast [387]. The DBT represents a more accurate diagnostic indicator than 2D imaging for evaluating the morphological features, e.g., shape and margin of the different immunophenotypes of the breast cancer, thus being able to play a crucial role in the molecular imaging and prognosis [388–392].

Over the last decade, DL has emerged as a promising computational approach for the automatic detection, classification, and segmentation of cancerous masses thorough the analysis of diagnostic medical images, thus enabling the CAD and clinical decision support systems [8, 329, 393, 394]. The DL methods along with the traditional image processing techniques have already been established as an effective approach to automatically analyze diagnostic images for breast cancer diagnosis and monitoring [7, 9, 10]. Numerous studies dealt with automatic detection, segmentation, and classification of the breast lesions that achieved considerably moderate to high performances [395–404].

However, the automatic classification of the breast lesions according to shape, size, and physical appearance remains a challenging task due to the varying shape that refers to different type and stage of the cancer [405] (see Figure 6.2). The breast cancer is morphologically

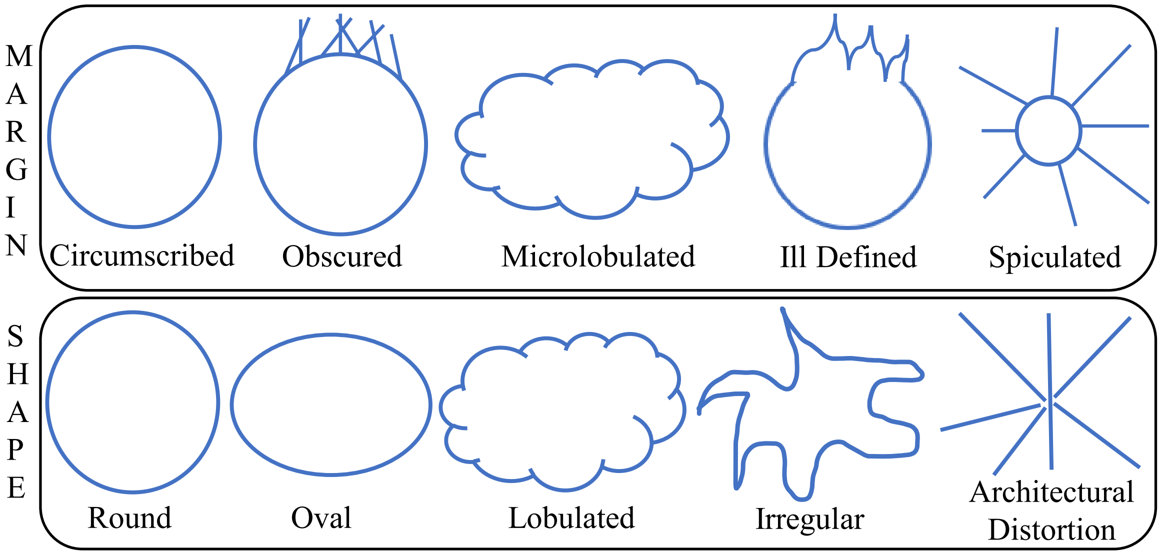


Fig. 6.2 The morphological division of the breast cancer shapes according to the growth pattern [40].

categorized into several varying shapes based on cancer growth pattern, named as round, oval, lobulated, irregular, and architectural distortion [40, 41].

Numerous existing studies deal with the shape-based breast cancer classification [405–407], however, most of these consider the mammogram instead of the DBT that offers several advantages as discussed above. A deep discussion of the state of the art is presented in the subsection 6.3.

6.3 Related Studies

Over the last decade, because of the superior aptitude to capture cancers, the DBT has become the new gold standard for the digital mastography [408]. Alongside this, machine learning has revolutionized the medical field by offering automatic detection, segmentation, and classification of the cancer [403, 404, 409–412].

The shape of the breast tumors leads to diagnosis of the different types and stages of the cancer [405]. Generally, the breast cancer is morphologically categorized into five shapes based on tumor growth pattern, named round, oval, lobulated, irregular, and stellar [40], as depicted in Figure 6.2. Numerous authors claim that the transition from the round shape to stellar shape of the cancer is the journey from benign to malignant cancer [405, 406, 413].

The shape-based breast cancer classification of mammogram images at RoI level using GAN and CNN is presented by Singh et al. [406]. The authors used a publicly available dataset for validation and achieved an overall classification accuracy of 80% for irregular, lobular, oval, and round shape classes.

Similarly, Kisilev et al. [41] proposed a multi-task loss CNN architecture based on the Faster R-CNN model to detect tumor lesions by considering irregular, round, and oval shapes of the breast cancerous lesions using in-house and publicly available datasets. Their approach generated bounding boxes around the tumor, and then used the semantic descriptors to identify the lesion shape inside the RoI. The accuracy on in-house and public datasets reached 88% and 82%, respectively, where both accuracy values were computed on accurately labeled data for testing purposes.

In a previous study by authors [384], two different approaches for the classification of DBT images into four lesions, i.e., irregular opacity, regular opacity, stellar opacity lesions, and no lesions, were implemented and tested on an in-house dataset. The first approach utilizes an artificial neural network that takes morphological and hand-crafted features extracted from the RoI images and performs classification. The second framework encompasses the pretrained CNNs without requiring the hand-crafted features. The authors claimed that the VGG network outperformed the other pretrained architectures by reaching 91.61% and 81.49% accuracy with and without augmentation.

A GAN-based interpretable CAD system for the classification of oval, round, irregular, and lobular shapes on the mammogram images was devised by Kim et al. [407]. The CAD system was tested on a public dataset that managed to achieve 71% accuracy on the lesion shape classification.

A study on mammogram and MR scans on three publicly available datasets was conducted by Shrivastava et al. [405] to classify the shapes of the tumorous regions using geometrical feature-based classifier. Since the authors merely considered the binary classification problem (benign lesion vs. malignant lesion), unlike previously explained methods, the reported accuracy, i.e., 91.4%, was pretty high.

A recent study by Sakai et al. used SVM, random forests, naive Bayes, and multilayer perceptron methods to classify the breast lesions on tomosynthesis images [414]. The authors also considered radiomic features along with the shape of the lesion. All the round and oval tumors were labeled benign, whereas the irregular and the stellar were labeled malignant on an in-house dataset. The best achieved accuracy value was 55% for round vs. oval classification, and 84% in case of irregular vs. stellar classification.

Said et al. [415] adopted the genetic algorithm to select the most significant hand-crafted features out of the total 130. Finally, the back-propagation neural network was employed for the classification task on round, oval, lobular, and irregular shapes that reached 84.5% accuracy on the digital database for screening mammography dataset.

6.4 Materials and Methods

6.4.1 Dataset

Back in 2016, a total number of 16 patients participated in breast tomosynthesis examination. The average age of all the considered subjects was 49.8 years with a standard deviation of 9.2 years. The patient with minimum age was 35, whereas the patient with maximum age was 65 years. Since few subjects underwent multiple trials, the total number of examinations summed up to 39.

This study inherits the RoI-level images generated in a previous study [384] aimed at constructing a dataset of RoIs that can be fed to the DL models for the shape-based classification, where the machine learning algorithms are employed to generate the tiles from the original images. Figure 6.1 shows the RoIs over the images after the segmentation phase, where in the case of None class (i.e., no lesion class), random images were taken from the area of the breasts containing no lesion.

A radiologist (University of Bari Medical School, Bari, Italy) with fifteen years of experience in the field of breast imaging labeled the images. In order to verify labeling accuracy, all radiological reports were assessed, including the histological reports for all detected lesions and 2 years' follow up with DBT for negative cases. The images were labeled and classified into four classes, comprising no lesions (None); irregular opacity (Ori); regular opacity (Oro); and stellar opacity (Ost). The None class contained 1000 images, whereas the Ori, Oro, and Ost classes contained 391, 654, and 480 lesion images, respectively, constituting a total number of 2525 samples.

6.4.2 CNN Models

In the subsection below, the CNN architectures considered for the classification task are briefly introduced.

- VGG

  The VGG [72] comes in two famous versions, with 16 and 19 layers comprising 144 million parameters. This study considers the earlier VGG-16, which consists of several number of channels, 3×3 receptive fields, and a stride of 1. This model is composed of convolution layers, max pooling layers, fully connected layers with 5 blocks and each block with a max pooling layer, and extra convolutional layers contained in the last three blocks.

- ResNet

  The deep neural networks suffer from the gradient vanish problem, which led to the development of Residual Network (ResNet) architecture. The ResNet takes care of the gradient vanishing problem and makes sure the performance remains satisfactory over the top and lower layers. ResNet comes with several variants where the number of layers is the distinguishing parameter among numerous architectures, however the underlying mechanism remains similar. This architecture utilizes skip connections between layers. The ResNet-34 and ResNet-50 [74] contain 34 and 50 layers and implement residual learning. This net is efficient to train and also improves the accuracy, which led to utilize the two versions of network for multiclass classification purposes in this work.

- ResNeXt

  The ResNeXt, a counterpart of ResNet, is a specifically designed image classification network with very few tuneable parameters. It contains a series of blocks with a set of aggregations of similar topology with an additional dimension called cardinality. This cardinality, which creates major difference between its brother networks, competes with the depth and width of the network [416]. The simpler architecture based on VGG and ResNet with fewer parameters yields better accuracy on ImageNet classification dataset. The word *NeXt* in the name of the network refers to next dimension which surpasses ResNet-101, ResNet-152, ResNet-200, Inception-v3, and ResNet-v2 on the ImageNet dataset in accuracy.

- DenseNet

  The DenseNet [417], or in other cases, dense convolutional network, is a type of CNN designed to guarantee the maximum information flow between all layers in the network. The layers are subjected to align the feature map size and connect among each other, forming a dense network. The DenseNet works on feed-forward principle. Each layer in the network receives the input from the preceding layer, grabs the additional input, and hands it over to the following layer along with the feature map. All the layers follow a similar analogy. Differently from ResNets, in which the features are not combined through summation before they are passed into a layer, the feature combination is performed by the concatenation of these ones.

DenseNet comes with several variants where the number of layers is the distinguishing parameter among numerous architectures, however the underlying mechanism remains similar. The DenseNet-121 and DenseNet-161 contain 121 and 161 layers and follow the feed-forward method. This net is efficient to train and also improves the accuracy, which led to utilize the two versions of network for multiclass classification purposes.

- SqueezeNet

  The SqueezeNet is another popular CNN model particularly known for its smaller size. The major motivations and reasons that caused this network to be smaller include the following: (a) during the procedure of training, the communication over the servers is shortened, (b) the minimum requirement of bandwidth for exporting a model from cloud to any other device is also cut, and (c) the smaller a model is, the less hardware and memory it requires to run.

  The SqueezeNet architecture is also simple; it contains 8 fire modules sandwiched between two convolutional layers. The sandwiched fire modules also contain a squeeze convolution layer with numerous filters of varying sizes. Each fire module comprises several filters that increase with respect to the network progression, being fewer in the start and more in the end. The SqueezeNet also utilizes the max pooling operation at several levels, including first and last layers.

  The SqueezeNet appears to achieve comparable accuracy to AlexNet on the ImageNet dataset with fifty-times-reduced number of parameters. It also offers scalability that implies that the size of SqueezeNet model can also be compressed to as low as 0.5 MB.

- MobileNet-v2

  The MobileNet-v2 [418], a depthwise separable convolutional network aimed at downsizing the model, is an architecture based on inverted residual connections. These residual connections appear between bottleneck layers. The total number of residual bottleneck layers in MobileNet-v2 count to 19 which follow the fully convolution layer comprising 32 filters. The network brings several benefits, including the time and memory savings with higher accuracy of results. The output of the model speaks to the validity of the architecture.

6.4.3 Experimental Workflow

Figure 6.3 shows the overall flow diagram of the experimental approach. As depicted, the experimental setup starts by fine-tuning the considered pretrained networks with three different datasets, i.e., the original one and two datasets obtained with two different data augmentation procedures. Thereafter, the features extracted by the features maps of all versions of fine-tuned and pretrained networks were analyzed with both t-SNE and UMAP. Finally, Grad-CAM and LIME were applied to the RoI images.

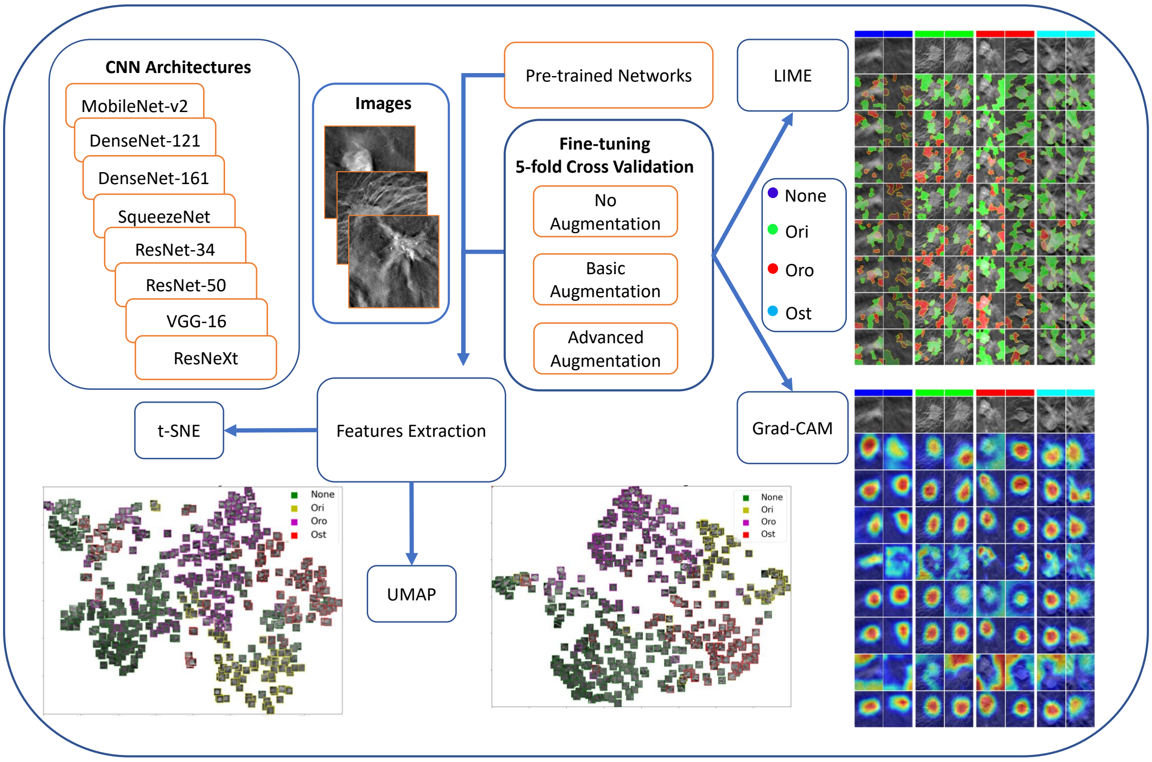


Fig. 6.3 The overall flow diagram of the experiments. The experimental setup starts by fine-tuning the considered pretrained networks with three different datasets, i.e., the original one and two datasets obtained with two different data augmentation procedures. Thereafter, the features extracted by the feature maps of all versions of fine-tuned and pretrained networks were analyzed with both t-SNE and UMAP. Finally, Grad-CAM and LIME were applied to the RoI images.

6.4.3.1 Data Augmentation Procedures

Due to the unavailability of large datasets, two types of augmentation were considered, i.e., basic and advanced. The basic augmentation comprises rotation and flip, whereas the advanced augmentation also includes color jittering. Numerous configurations with respect to data augmentation were considered, as reported in the Table 6.1 and described hereunder. By exploiting the *transforms.Compose* interface provided by PyTorch [68], the augmentations are sequentially performed on the fly, each with a given probability that has been set to 0.25.

No Aug refers to the adoption of no augmentation, with the exception of normalization, by rescaling intensity values of images from integers ranging $[0,255]$ to float values in $[0,1]$. **Basic Aug** consists of performing random rotation by the degrees in multiples of 90, and performing random horizontal and vertical flips. **Adv Aug** takes advantage of `ColorJitter` transformations in addition to the previous configuration of basic augmentation. The advanced augmentation comprises random perturbations of brightness, contrast, saturation, and hue. Finally, normalization is performed similar to **No Aug**.

6.4.3.2 Training Procedures and Cross-Validation

This study also implements the Transfer Learning (TL) paradigm using the weights of eight well-known CNN architectures, which not only saves the computational time but also produces higher performance outcomes. The major benefit of using TL comes into practice when the available dataset is not sufficiently large, whereas the performance also remains considerable on small datasets. For the classification problems, applying a pretrained model seems more rational rather than developing a model from scratch. This approach is also referred to as TL because the pretrained models' weights are transferred to other models to address the similar image classification problems.

Moreover, since the manual tuning of parameters is a time-consuming and less efficient process, this study encompasses the grid search to initially select, but later on settles to the learning rate of 1×10^{-5}, batch size of 32, and number of epochs to 50. Furthermore, a range of optimizers is available which can be selected depending upon the nature of problem; however, in this work, the Adam optimizer is used due to the simplicity and effectiveness on the classification problems. The used loss function was the cross entropy. Additionally, moving towards the train–test split, 5-fold cross-validation with stratification is performed in such a manner that approximately 80% of the data belonging to each class resides in train partition, whereas the remaining 20% dwells in the validation set.

6.4.4 Classification Performance Assessment

The results of all pretrained and fine-tuned nets are analyzed based on AUC. The mean and standard deviation of AUC are computed for each classifier among 5-fold results. The AUC

Table 6.1 Data augmentation summary. All augmentations are done on-the-fly with 0.25 probability in the order they are presented in the table. Normalization is always performed at the end after all other augmentations. `ColorJitter` refers to the random alterations of the *brightness*, range: $[0.8, 1.2]$; *contrast*, range: $[0.8, 1.2]$; *saturation*, range: $[0.8, 1.2]$; and *hue*, range: $[-0.2, 0.2]$

Transform	**No Aug**	**Basic Aug**	**Adv Aug**
`RandomRotation90`	✗	✓	✓
`RandomRotation180`	✗	✓	✓
`RandomRotation270`	✗	✓	✓
`RandomHorizontalFlip`	✗	✓	✓
`RandomVerticalFlip`	✗	✓	✓
`ColorJitter`	✗	✗	✓
`Normalization`	✓	✓	✓

and the standard deviation are also computed for each individual class against all architectures in three augmentation configurations.

Furthermore, the training and validation losses during the experimental procedure are also plotted to investigate the eventual problems that arise during the potential overfitting at each epoch. All the experiments are performed on a machine running on Windows 10 operating system, and a Python 3.7 environment is exploited with *PyTorch* (`torch` v1.10.0, `torchvision` v0.11.0), `grad-cam` v1.3.6, and `lime` v0.2.0 libraries for DL and XAI. To this end, CUDA 11.3 is used to take advantage of the GPU power.

6.5 Experimental Outcomes

The section below illustrates the experimental results of the study in terms of the classification performance, XAI outcomes, and the relevant training and validation trends. The section contains a comparative analysis of the employed techniques and highlights the identified significant trade-offs.

6.5.1 Performance Module

The summary of the experimental results of all eight CNN models considered in this study in terms of AUC with 5-fold cross-validation is provided in Table 6.2 for the three conceived experimental configurations, i.e., without augmentation (*No Aug*), with basic augmentation (*Basic Aug*), and with advanced augmentation (*Adv Aug*), respectively.

Table 6.2 The summary of the results obtained for **No Aug**, **Basic Aug**, and **Adv Aug** configurations is provided hereunder. The bold text represents the best value of the corresponding parameter among all CNN models, that is mean over all four classes

Architecture	**Area Under the Curve (AUC)**		
	No Aug (None, Ori, Oro, Ost)	**Basic Aug** (None, Ori, Oro, Ost)	**Adv Aug** (None, Ori, Oro, Ost)
MobileNet-v2	91.9 ± 1.1	92.4 ± 0.9	93.6 ± 1.2
	97.4 ± 0.4	98.0 ± 0.6	97.6 ± 0.9
	95.2 ± 1.3	95.9 ± 1.1	96.3 ± 0.9
	95.8 ± 0.7	96.6 ± 0.5	96.5 ± 0.7
	95.1	95.7	96.0
DenseNet-121	90.1 ± 1.2	93.9 ± 1.9	94.5 ± 1.3
	94.2 ± 1.4	98.5 ± 0.6	98.2 ± 0.8
	89.9 ± 1.7	95.5 ± 0.6	96.7 ± 0.8
	92.9 ± 1.8	97.1 ± 0.8	97.2 ± 1.2
	91.8	96.2	96.6
DenseNet-161	94.8 ± 0.9	95.8 ± 1.0	96.4 ± 0.5
	97.6 ± 1.4	99.1 ± 0.7	99.4 ± 0.2
	95.8 ± 1.3	97.8 ± 1.0	98.7 ± 0.7
	97.0 ± 0.9	98.2 ± 0.3	98.0 ± 0.7
	96.3	97.7	**98.2**
SqueezeNet	50.9 ± 3.0	56.6 ± 5.6	62.7 ± 8.1
	85.9 ± 3.2	84.3 ± 1.4	86.4 ± 2.9
	68.9 ± 5.6	67.6 ± 3.8	71.7 ± 7.2
	83.8 ± 2.6	86.2 ± 3.7	87.6 ± 3.1
	72.4	73.7	77.1
ResNet-34	92.0 ± 0.8	94.5 ± 1.0	95.4 ± 0.6
	96.2 ± 0.8	98.6 ± 0.5	98.9 ± 0.5
	94.7 ± 1.7	97.6 ± 0.4	97.4 ± 1.0
	96.1 ± 1.3	97.6 ± 0.7	97.7 ± 0.7
	94.8	97.1	97.3
ResNet-50	93.8 ± 1.1	95.3 ± 1.2	96.2 ± 0.6
	98.0 ± 0.5	99.4 ± 0.3	99.3 ± 0.3
	95.8 ± 0.8	97.8 ± 0.6	97.9 ± 0.7
	97.0 ± 1.0	97.8 ± 0.9	98.5 ± 0.4
	96.1	97.6	98.0

Architecture	Area Under the Curve (AUC)		
	No Aug (None, Ori, Oro, Ost)	**Basic Aug** (None, Ori, Oro, Ost)	**Adv Aug** (None, Ori, Oro, Ost)
VGG-16	90.6 ± 1.7	92.5 ± 1.4	93.6 ± 1.3
	98.1 ± 0.6	98.9 ± 0.6	97.7 ± 0.7
	96.1 ± 0.7	96.7 ± 0.7	97.2 ± 0.7
	96.6 ± 0.4	97.7 ± 0.6	98.1 ± 0.6
	95.3	96.4	96.6
ResNeXt	94.1 ± 1.0	96.1 ± 0.7	95.8 ± 0.7
	97.7 ± 0.7	99.3 ± 0.2	99.0 ± 0.7
	96.0 ± 0.8	97.9 ± 0.7	98.2 ± 0.8
	97.1 ± 0.5	98.3 ± 0.3	98.2 ± 0.9
	96.2	**97.9**	97.8

6.5.1.1 Classification Results

In the case of *No Aug* configuration, it can be observed from Table 6.2 that DenseNet-161 is the architecture with the highest mean AUC of 96.3%. The ResNeXt and ResNet-50 networks are slightly behind, with AUC of 96.2% and 96.1%, respectively. The MobileNet-v2, ResNet-34, and VGG-16 collectively form a third cluster with AUC of around 95%. Conversely, the SqueezeNet is the worst-performing model in this experimental setup, managing to achieve merely 72.4% AUC.

In the case of the *Basic Aug* configuration, all architectures performed considerably better than the previous *No Aug* configuration. The results reveal that ResNeXt obtained the highest AUC of 97.9%, beating all other architectures. The DenseNet-161 and ResNet-50 achieved similar performances with the AUC of 97.7% and 97.6%, respectively. Once again, the performance of the SqueezeNet failed to present significant outcomes, thus abiding by the *No Aug* configuration.

The second augmentation setup, called *Adv Aug*, emerged to be even better than both previously conceived *No Aug* and *Basic Aug* setups. The DenseNet-161 reached the top AUC of 98.2%. The ResNet-50 appeared to be the second best model, with a slightly lower AUC of 98.0%.

Finally, as noted during the *No Aug* and *Basic Aug* configurations, the SqueezeNet is the model which offers least reliability with the largest inter-fold variability; however, it improved the AUC from the previous setups.

Therefore, it can be summed up that the ResNeXt and DenseNet-161 remain the top-performing models, and the augmentation configurations considerably improved the performance of all CNN architectures. However, the SqueezeNet failed to produce convincing results.

6.5.1.2 Train and Validation Loss Trends

The training and validation losses fluctuate with respect to each epoch. All models were run at different values of epoch starting from 10 up to 50; however, for the purpose of clarity and concision, only the results obtained considering the 50 epochs are illustrated.

The loss curves demonstrate important trends to monitor in order to clearly distinguish the working mechanism of the CNN architectures over the repeated iterations. In Figure 6.4, it is distinctive to visualize the loss on both train and validation sets (first fold) for the best, i.e., DenseNet-161, and the worst, i.e., SqueezeNet, CNN architectures in the case of *No Aug* configuration.

Although the SqueezeNet shows decreasing loss on both train and validation sets in Figure 6.4b, the training loss curve becomes constant right after fewer epochs in DenseNet-161 in Figure 6.4a. Moreover, the validation curve depicts increasing behavior after fewer than ten epochs for the DenseNet-161. Such behavior could be motivated by the huge number of parameters that might cause the overfitting problem on the train set.

The train and validation loss curves considering the advanced data augmentation configuration are provided in Figure 6.5. The augmentation helped the DenseNet-161 to overcome the increasing validation loss, as shown in Figure 6.5a. This evidences that incorporating on-the-fly data augmentation solved the overfitting issues. However, the SqueezeNet struggles to keep the loss low, as depicted in Figure 6.5b, and ends up with even worse performance than the no augmentation configuration. Differently from DenseNet-161, the SqueezeNet does not seem to take advantage of the on-the-fly augmentation, possibly due to lower number of parameters.

Additionally, the reported behavior of the loss trends on both train and validation sets is comparable to the other folds. With the intention of concision, only the outcomes of the best and the worst performing architectures are depicted, i.e., DenseNet-161 and SqueezeNet, respectively, in terms of AUC and loss.

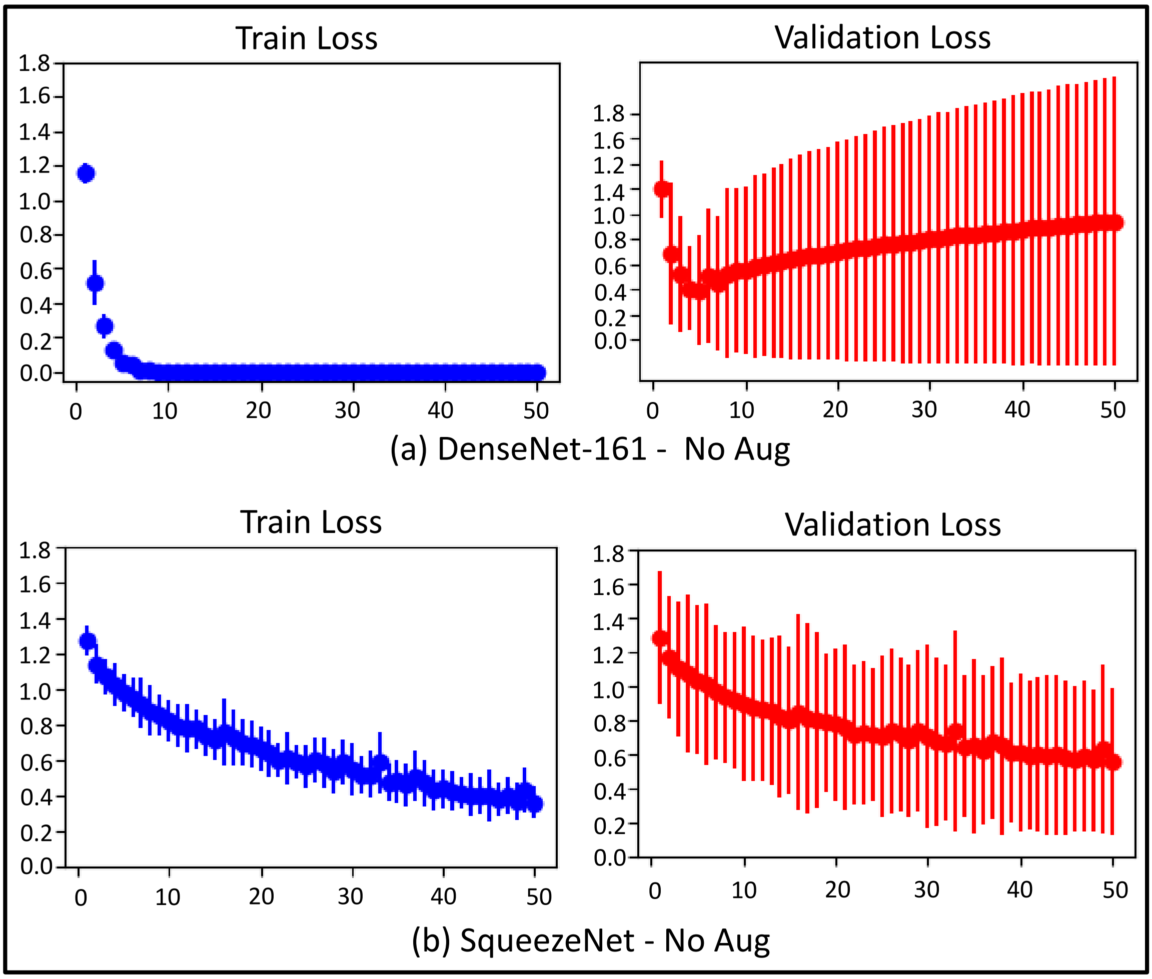


Fig. 6.4 The train and validation loss for the fold = 0 of cross-validation. (**a**) DenseNet-161 with *No Aug* configuration; (**b**) SqueezeNet with *No Aug* configuration. The reported behavior of train and validation loss trends is comparable to that of the other folds.

6.5.2 Area under the Curve and Number of Parameters Trade-Off

During the experimental phase, the author came across an interesting trend between the mean AUC (computed on the test set) and the number of parameters of the employed architectures. A plot illustrating the relationship between AUC and the number of parameters for the eight considered CNNs is presented in Figure 6.6. It is observable that the VGG-16 holds a gigantic number of parameters but without yielding the corresponding improvement in the AUC. The SqueezeNet, on the contrary, is a small architecture in terms of number of parameters, but fails to realize commendable AUC among the contemplated models. The best trade-off between the number of parameters and the performance can be seen in ResNet-like models, with ResNet-50 winning the dispute.

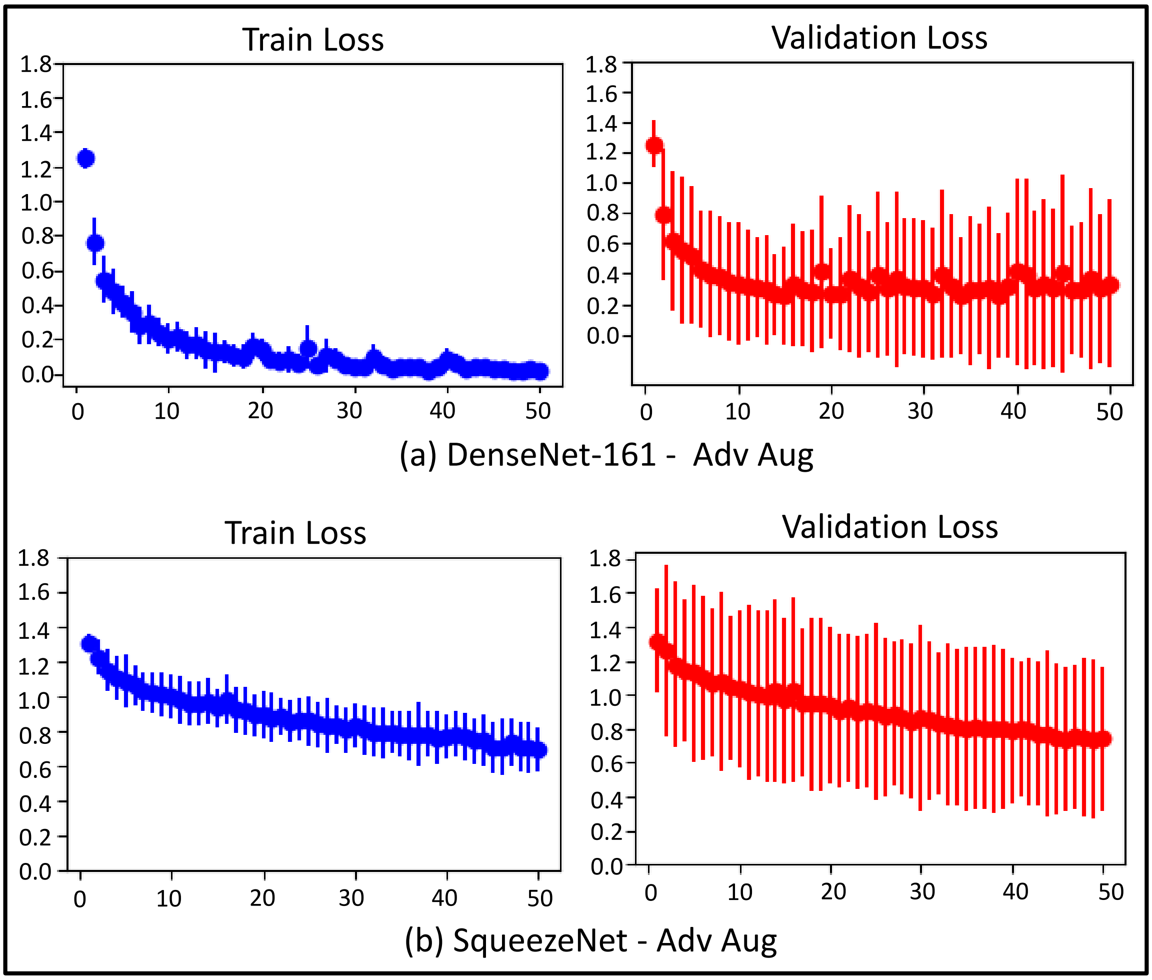


Fig. 6.5 The train and validation loss for the fold = 0 of cross-validation. Figure (**a**) DenseNet-161 with *Adv Aug* configuration; (**b**) SqueezeNet with *Adv Aug* configuration. The reported behavior of train and validation loss trends is comparable to that of the other folds.

6.6 Discussion

This study proposes a novel, visually explainable DL-driven multiclass shape-based breast cancer classification framework for tomosynthesis lesion images. For the task of morphological classification, eight DL models are employed on tomosynthesis breast images and two families of XAI methods, i.e., perceptive interpretability and mathematical interpretability, are incorporated to explain the results acquired during the validation study in order to create the trust among the clinicians and AI. The relevant material to the XAI and the results are described in the next chapter of the thesis.

The CAD system developed in this study is able to encircle the potential growth pattern of the tumorous regions on the DBT images and results in the improved diagnostic and

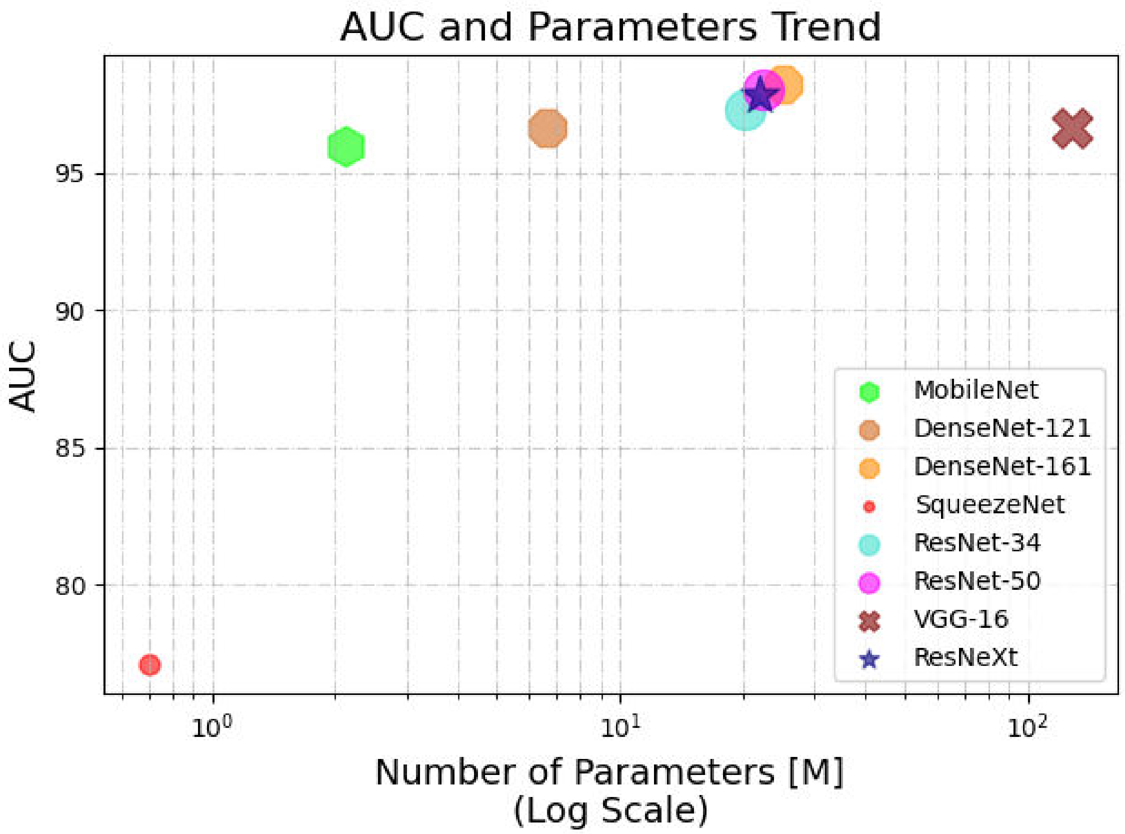


Fig. 6.6 The relationship between area under the curve and number of trainable parameters for the eight CNN architectures considered throughout this study.

prognostic performance. The successful implementation also enhances the trustworthiness among the clinical field and the high-accuracy-yielding DL architectures. The subsection below comparatively discusses the shape-based breast cancer classification.

6.6.1 Shape-based Breast Cancer Classification

Quantitatively, the extensive experimental results are elaborated, considering the pretrained DL methods on both with and without data augmentation configurations. The mean AUC values of the developed models improved during the augmentation phases. The crown of overall best performing algorithm belongs to DenseNet-161 due to persistent performance, i.e., reaching higher than 96.0% across *No Aug*, *Basic Aug*, and *Adv Aug* setups.

In particular, the best-performing model, i.e., DenseNet-161, increases by 1.45% and 1.97% in the mean AUC from *No Aug* to *Basic Aug* and *Adv Aug*, respectively. It impressively increases by 33.01%, 33.28%, and 27.10% over the SqueezeNet in configuration-to-configuration comparison, i.e., *No Aug* to *No Aug*, and so on. In the case of *Basic Aug*, the ResNeXt outperforms all other architectures with a percent increase of 33.56 from the worst-performing model.

Since the results are comparable, any particular model performing best in terms of AUC and loss may not perform ideally in all aspects. The reason is the primitive learning and weight updating mechanism of the CNN models. For example, in the *No Aug* phase, three out of four individual AUC values of ResNeXt among classes remain higher than the respective individual AUC values of the DenseNet-161, despite the equal mean AUC.

The utilization of the augmentation techniques revamped the trends of the validation loss, as shown in Figures 6.4 and 6.5; however, the improvement in the validation loss is negative for the worst-performing model, i.e., SqueezeNet, which fluctuates between 0.3 to 1.3 and 0.7 to 1.3 for *No Aug* and *Adv Aug* configurations, respectively. Both the training and validation losses increased.

The illustrated loss values are evidently coherent to the fact that a huge model such as DenseNet-161, with tons of parameters, overfits when it is trained with no augmentation over an increasing number of epochs in this experimental setup. Instead, SqueezeNet has the opposite problem, being unable to even properly comprehend the fundamental patterns, resulting in an underfit behavior. After augmentation, underfitting problem of SqueezeNet cannot be resolved, as shown by comparing Figures 6.4b and 6.5b, but the overfitting issue of the DenseNet-161 is mitigated as presented by comparing Figures 6.4a and 6.5a.

A noteworthy consideration arises when considering the performance of a model in relation to its size and complexity. In Figure 6.6, a noteworthy trend exists between the number of parameters and the AUC, the models having a huge number of parameters compromised at the mean AUC at certain levels. On the contrary, the models with an extremely low number of parameters may result in bad generalization performance, since with reduced number of parameters, the model is hardly able to learn simple patterns in this study.

Nevertheless, the two different augmentation configurations and three different execution setups (i.e., 10, 30, and 50 epochs) disclose a clear improvement with augmentation in DBT classification framework. The basic augmentation improves performance compared to no augmentation, and the advanced augmentation plays its part and further increases the AUC outcomes. One major reason reckoned is the considerably high visually noticeable resemblance between the training and the validation data, whereas the state-of-the-art architectures, the significant clinical data, and the RoI-level cropped images may possibly be other driving causes. Finally, from several above discussed dimensions, the study proves the applicability of the CNN models in the classification task of DBT lesions on RoI level. Taking the advantage of TL, the framework reaches efficient results with fine-tuning of several parameters, and paves the way towards autonomous CAD systems.

6.7 Limitations

In this work, the pretrained DL models are employed for the classification tasks using a 5-fold cross-validation strategy. Since the train and validation data come from the same source, few models suffer from generalizability and overfitting problems. The models can be validated on the external datasets after training for better understanding. Additionally, the dataset used in the study is relatively small; this is a major reason to incorporate the pretrained models and perform fine-tuning also with data augmentation. The used models could be trained and tested with large-scale datasets acquired from different cohorts.

6.8 Summary

Breast cancer is the leading deadly ailment in women, and its inevitable progression has become a major concern for the healthcare industry. However, timely diagnosis can significantly improve the medication and prevent the further expansion of the cancerous regions. DL offers great success in automatic detection and classification using medical imaging data. However, the black-box nature of the decision-making mechanism of the DL architectures hampers the trust among the clinicians. The XAI techniques uncover the black-box and hidden nature of the DL and provide useful apprehension of the high-accuracy-yielding DL models. This builds confidence in machine learning in the clinical domain and paves the way towards DL-centered image-guided CAD systems.

In this work, a robust visually and mathematically explainable DL framework for multi-class shape classification of tomosynthesis breast lesion using eight pretrained CNN models using an in-house dataset is proposed. Due to small-scale data availability, the data augmentation was incorporated. The best fine-tuned model achieved mean AUC values of 98.2% and 96.3% with and without considering the data augmentation, respectively.

Chapter 7

Explainable Artificial Intelligence

7.1 XAI for DL in Medical Imaging

In spite of the enormous success, the complex nature of the DL techniques hides any possible information of the underlying decision mechanism [31, 32], which questions its usage in the healthcare domain where explainability holds paramount significance to build a trust on decisions made by surging AI. XAI brings forward the possibility of explaining the results of DL models and reveals how the models produce these results. Generally, XAI is supposed to fit a model onto four basic attributes [33]:

- *Transparent*: open to the degree where humans can understand the decision-making mechanism.
- *Justifiable*: the decision can be supported or justified along each step.
- *Informative*: to provide reasoning and allow reasoning.
- *Uncertainty yielding*: does not follow hard-coded structure, but open to change.

XAI has drawn a tremendous amount of attention in the recent past (see Figure 7.1) and it is not hard to comprehend the importance of such methodologies in the clinical field, where AI is expanding enormously [34]. Such new research topic is extremely fascinating yet challenging, because as it can be easily envisaged, a more complex AI model that can reach high-level performance is less interpretable than, for example, a simple rule-based model (see Figure 7.2).

Numerous XAI methods and relative updated versions have been proposed in the literature [34]. The presented approaches can be classified into two major categories: perceptive

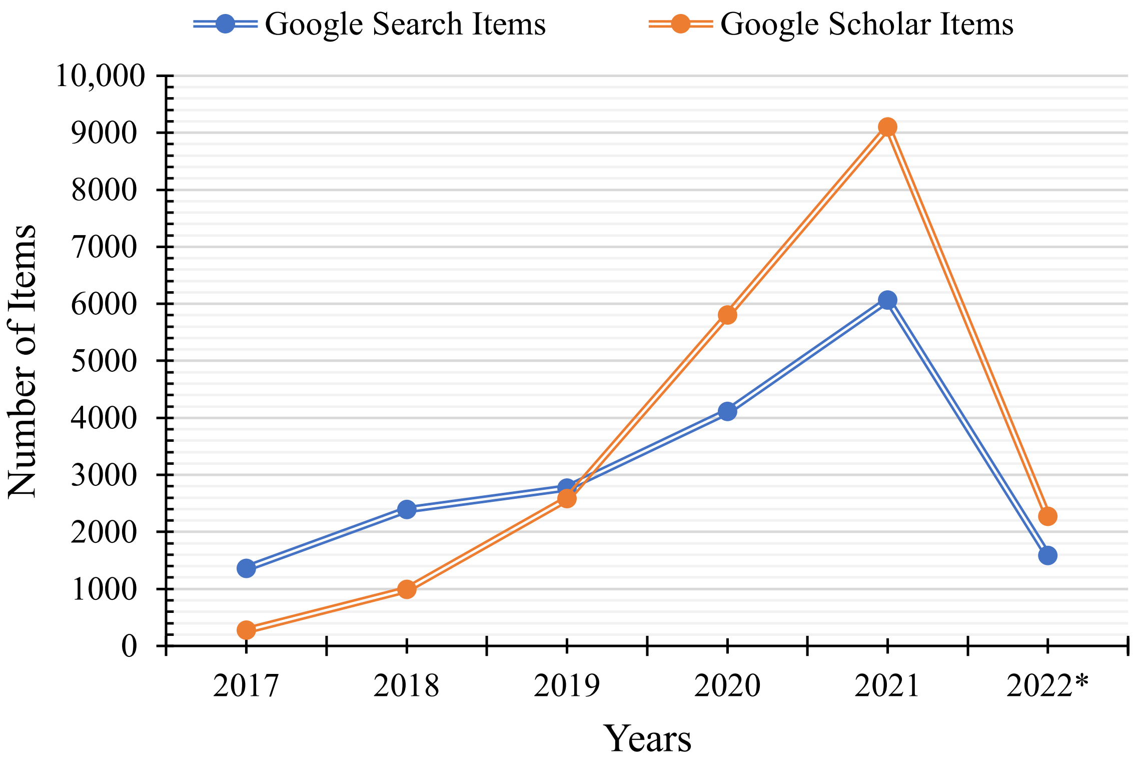


Fig. 7.1 The popularity index of the term *'explainable AI'* over the period of 2017–2022. Google Item Search indicates the queries in Google search engine, whereas Google Scholar Search points out the published studies available at Google Scholar (* results until March 2022 are extracted).

interpretability and mathematical interpretability. The former includes visual interpretability that can be visually perceived by humans, for example, the heatmaps that report the importance of input and their contribution to that decision. The mathematical-interpretability-based methods usually rely on simple models, e.g., linear models, or on correlation/clustering methods that analyze the extracted features. When visual evidence is not useful or erroneous, the mathematical evidence can also be used as a complement for the interpretability. Therefore, various methods should be applied simultaneously in order to provide reliable interpretability [419].

In this part of the study, the XAI methods are applied to evaluate the results of DL models on the previously proposed breast cancer classification pipeline presented in the Chapter 6. The developed framework is a CNN based DL framework for the classification of lesions according to the shape pattern by analyzing the RoI on DBT images.

The trained DL models and related results have been further interpreted by incorporating two different methodologies for each of the two explanation mechanisms. Gradient-weighted

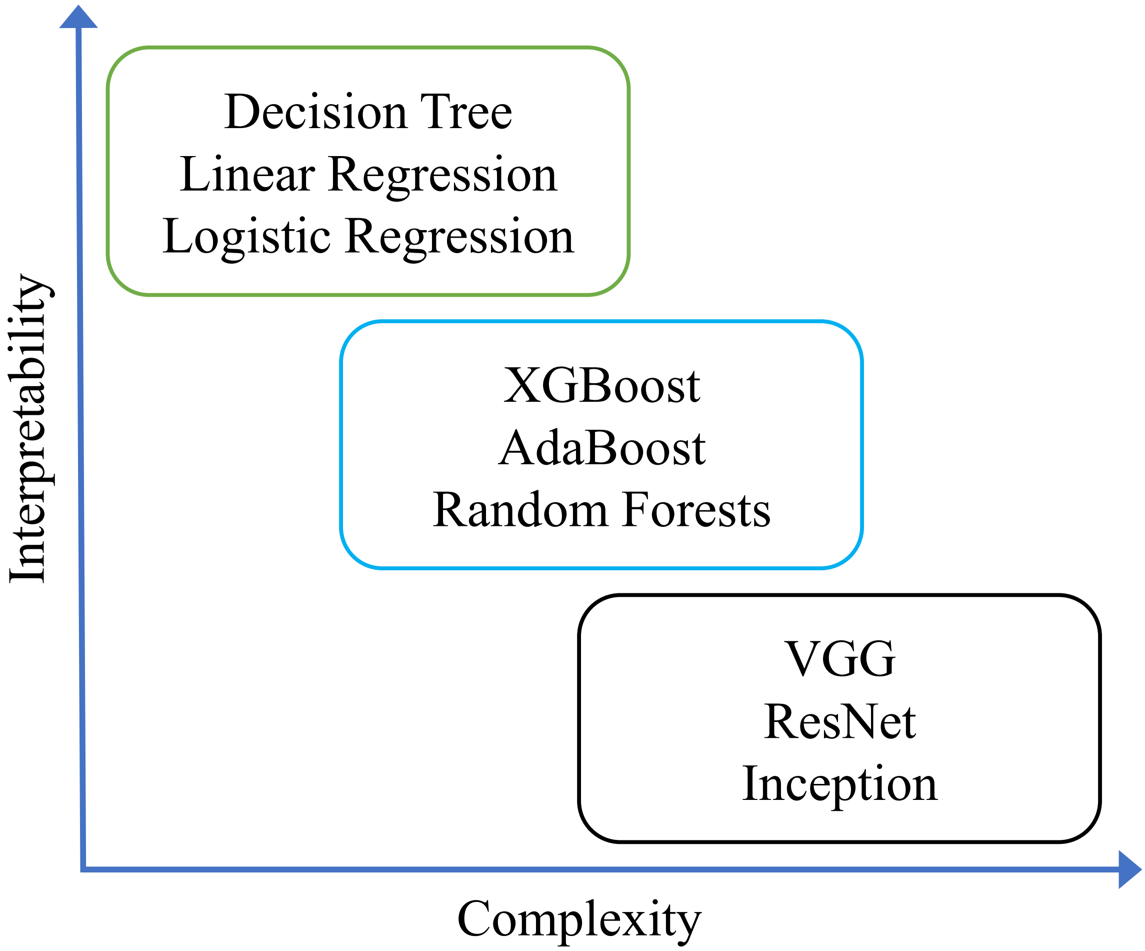


Fig. 7.2 The complex models are less explainable as compared to the simple models, because of the increasing number of hidden layers and parameters. The more simple a model is, the more interpretable it is.

Class Activation Mapping (Grad-CAM) method and Local Interpretable Model-agnostic Explanations (LIME) have been used to visually interpret the results, whereas t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) techniques have been utilized to study the mathematical interpretability of the features automatically extracted by all eight CNN architectures [34].

7.2 XAI and Breast Cancer

Several studies in the literature dealt with the breast cancer imaging and XAI. Ricciardi et al. proposed a binary classification framework based on AlexNet and VGG-19 architectures to recognize the presence or absence of mass lesion on DBT image of two in-house datasets [420]. The authors adopted the Grad-CAM method to study the behavior of the classifiers, i.e., whether they align with the delineated lesion labeled by the expert radiologists. Employing the Grad-CAM method, the authors concluded that central areas of the lesion contribute more towards classification, whereas the branches of the tumor bring less impact on classification.

Masud et al. performed multiclass classification of ultrasound breast images considering benign, malignant, and normal classes on two public datasets [421] using eight pretrained CNNs and a custom model. The highest accuracy among the pretrained architectures was achieved by ResNet-50 with a value of 92%, whereas the customized model achieved 100% accuracy. For explaining the classification mechanism and to study the performance of the customized model, the Grad-CAM heat map visualization was also incorporated.

Suh et al. compared the binary classification performance of DenseNet-169 and EfficientNet-B5 models on predicting the availability of malignancy of the lesions from the mammogram images on an in-house dataset [422]. The former network achieved an accuracy of 88.1%, whereas the latter reached to 87.9%. The Grad-CAM method was used merely to spotlight the important regions over an image that lead to the classification. The authors claimed that Grad-CAM also spotlights the surrounding areas of the tumor, which shows the importance of not only the tumorous region but also the nearby regions.

Similarly, Lou et al. [423] proposed a framework driven by a custom model and a pretrained (ResNet-50) architecture to classify the benign and malignant masses on two publicly available mammogram datasets. The authors reached an accuracy of 83.75%. The Grad-CAM is employed to examine the spatial position of the object located by the CNNs. The authors claim that in case of successful classification, the XAI method highlights the mass correctly, however, it may also focus on the irrelevant regions due to spots that are not lesions.

Apart from the unavailability of explainability in majority of the existing articles dealing with DBT image classification task, only few authors [41, 406, 407, 415] considered the shape-based cancer classification of the lesion, which not only distinguishes among the normal and abnormal images but also highlights the growth pattern of the tumor shapes.

Unlike the proposed multi-class morphological CAD classification framework in this study, most of the authors merely focused on malignant vs. benign classification of the lesion [405, 409–412, 414] and provided no, or unsatisfactory, XAI discussion in some cases. The only two authors which provided XAI in their CAD system [420, 422], limited it to the Grad-CAM method, and did not consider more complex classifications of the breast cancer, such as the shape one investigated by this study.

The main contributions of the study can be stated as (i) to investigate the applicability of both perceptive and mathematical XAI methods at RoI level in the DBT images; (ii) to investigate the reliability of features and learning process and correlate it with the overall DL model performance; (iii) to perform a comprehensive comparison of the CNN architectures and the XAI methods in order to guide the engineers and the radiologists interested in implementing DL-driven CAD systems.

7.3 Materials and Methods

The interpretability and explainability have largely been achieved by applying two families of methods, namely, perceptive interpretability and mathematical interpretability [34]. The perceptive XAI is responsible for bringing a straightforward view of the top contributing features that affect the final predictions, whereas the mathematical interpretability provides insights into the used models and portrays the features that are employed to make the final predictions. The former is used to study the feature-level classification behavior (the importance of a particular region towards classification) of the DL architectures, whereas the latter is used to study the clustering capabilities of the networks.

7.3.1 Perceptive XAI

This study adopts two of the most widely admired XAI-based perceptive explanation methods called Grad-CAM and LIME [34] in order to explain the decisions made by the CNN architectures. Both the models are post hoc (i.e., they take as input an already trained model [34]) and can be extended to any DL network for explanation without any alteration in the rudimentary mechanism of the DL methods. Below, a brief description of the Grad-CAM and the LIME models is reported.

7.3.1.1 Grad-CAM

According to Das et al. [424], Grad-CAM can be classified as a back-propagation-based method, meaning that the algorithm makes several forward-passes (one or more) through the neural network and generates attributions during the back-propagation stage using partial derivatives of the activations. Contrary to the CAM, which requires a particular pattern of network under analysis, the Grad-CAM is the generalization that can be applied without any modifications in the DL model [425].

The Grad-CAM produces a heatmap of the class activation in response to the input image and a class. In other words, for a particular provided class, the Grad-CAM produces approximate and comprehensible representations of the network's decision-making mechanism in the form of a heatmap that translates to the feature importance. Specifically, in the last layers of a CNN, neurons look for semantic information associated with a specific class. In this layer, Grad-CAM uses the gradient flowing into it to assign a weight to each neuron according to its contribution to the decision in the classification task. The computed information is translated into a jet color scheme to depict the saliency zones, where the red color represents the higher intensity, i.e., pixels on which the network is focusing more for performing the classification, while the blue color represents the lower intensity of the focus. The neuron-importance weights can be computed as inscribed below [425]:

$$w_k^c = \frac{1}{S}\sum_i \sum_j \frac{\partial y^c}{\partial M_{ij}^k} \tag{7.1}$$

Where:

- $\frac{1}{S}\sum_i \sum_j$ is the global average pooling (i, j are respectively the indexes of width and height dimensions and S is the total number of cells in a feature map).
- y^c is the activation class score for target class c.
- $\frac{\partial y^c}{\partial M_{ij}^k}$ is the gradient computed via backpropagation, for a target class c, with D_{ij}^k as the activation of cell at spatial location i, j for a feature map M^k.

Then the weighted combination of forward activation maps is performed, followed by a ReLU, obtaining the *class-discriminative localization map Grad-CAM*:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k w_k^c D^k\right) \tag{7.2}$$

7.3.1.2 Local Interpretable Model-Agnostic Explanations

In this study, another well-known explanation technique based on model-agnostic phenomena known as LIME is incorporated. The approach can be applied to any DL model. Specifically:

- *Local*: states that LIME explains the behavior of the model by approximating its local behavior;
- *Interpretable*: emphasizes the ability of the LIME to provide an output useful to understand the behavior of the model from a human point of view;

- *Model-Agnostic*: means that LIME is not dependent on the model used; all models are treated as a black-box.

In this classification problem, the explanation of LIME remains simple. It takes the superpixels (a patch of pixels) of the original input image after generating a linear model, and generates several samples by exploiting the superpixels. The quick-shift algorithm is responsible for the computation of superpixels of an image. Thereafter, the perturbation images are generated and the final prediction is made.

Afterwards, a heatmap appears over the image that highlights the important pixels, i.e., regions that contribute in classification. The positively contributing features are highlighted in green while the negatively contributing superpixels are colored in red. The LIME also allows to pick a threshold value to select the number of top contributing pixels, either positively or negatively.

7.3.2 Mathematically Explained XAI

This section introduces two widely adopted and useful techniques for performing the task of mathematical interpretability implemented in the presented work. The mathematical interpretability offers t-SNE and UMAP techniques to represent the high-dimensional graph into lower dimensional space without compromising on the clustering structure.

Primarily, both the t-SNE and UMAP are meant for visualization; however, the main difference lies in the interpretation of the distance between the clusters. The t-SNE merely preserves the local structure in the data, whereas the UMAP can preserve both local and global structure in the data, which means that unlike the UMAP, the dissimilarity and the distance between clusters can not be interpreted with the t-SNE.

7.3.2.1 T-Distributed Stochastic Neighbor Embedding

The t-SNE [426] is a variation of the SNE technique that makes the visualization of high-dimensional data possible by associating with each datum a location in lower dimensional space of two or three dimensions. It has been developed to face two issues that affect SNE technique:

1. The optimization of the cost function, by using a variation of SNE cost function (symmetrized) and using a Student's t distribution for the computation of similarity between two datapoints in the lower-dimensional space.

2. The so-called “crowding problem”, by using a heavy-tailored distribution in low-dimensional space.

7.3.2.2 Uniform Manifold Approximation and Projection

The UMAP [427] is a nonlinear technique for the dimensionality reduction. It is based on three assumptions:

1. Data are uniformly distributed on an existing manifold;
2. Topological structure of the manifold should be preserved;
3. Manifold is locally connected.

The UMAP method can be divided into two main phases: learning a manifold structure in a high-dimensional space and finding the relative representation in the low-dimensional space. In the first phase, the initial step is to find the nearest neighbors for all datapoints, using the nearest-neighbor-descent algorithm.

Then, UMAP constructs a graph by connecting the neighbors identified previously; it should be noticed that the data are uniformly distributed across the manifold, so the space between datapoints varies according to regions where data are denser or sparse. According to this assumption, it is possible to introduce the concept of *’edge weights’*: from each point, the distance with respect to the nearest neighbors is computed, so the edge weights between datapoints are computed, but there exists a problem of disagreeing edges.

7.3.3 Performance Assessment Module

At this stage, the saliency maps using Grad-CAM algorithm are analyzed, and superpixel importance (both positive and negative) with the LIME technique is used to inquire what aspects the classifiers are focusing on, so as to build a trust for CAD systems that can be exploited to support the radiologists’ diagnostic workflow are computed. Generally, these methods identify which features oblige a DL model to discriminate among different lesions present on the image.

Particularly, to generate the heatmap visualizations from the Grad-CAM, all the architectures, except the VGG-16, utilize the last layer before the global average pooling layer. In case of VGG-16 architecture, the Grad-CAM is run at the maxpool layer before the first fully connected layer. Note that VGG16 is the only CNN among the considered architectures in this study that does not implement global average pooling, since it is an old architecture based on stack of fully connected layers at the end.

On the other hand, the LIME is a model-agnostic method; therefore, it creates the perturbations once the CNN finishes the classification task. In both cases, the specific class modeled as base class differs for all the networks; therefore, the labeled and the targeted classes are provided within figures. Moreover, the t-SNE and UMAP embeddings, before and after the fine-tuning of all architectures on the DBT image training set, are computed to understand how well TL approaches work on the radiological image scenario. The feature sets considered for t-SNE and UMAP are the same as for Grad-CAM discussed above.

7.4 Experimental Outcomes

This study employed XAI techniques from two families comprising mathematical interpretability, i.e., t-SNE and UMAP, and perceptive interpretability, i.e., Grad-CAM and LIME. The experimental outcomes of both XAI approaches on all CNN models are explained hereunder.

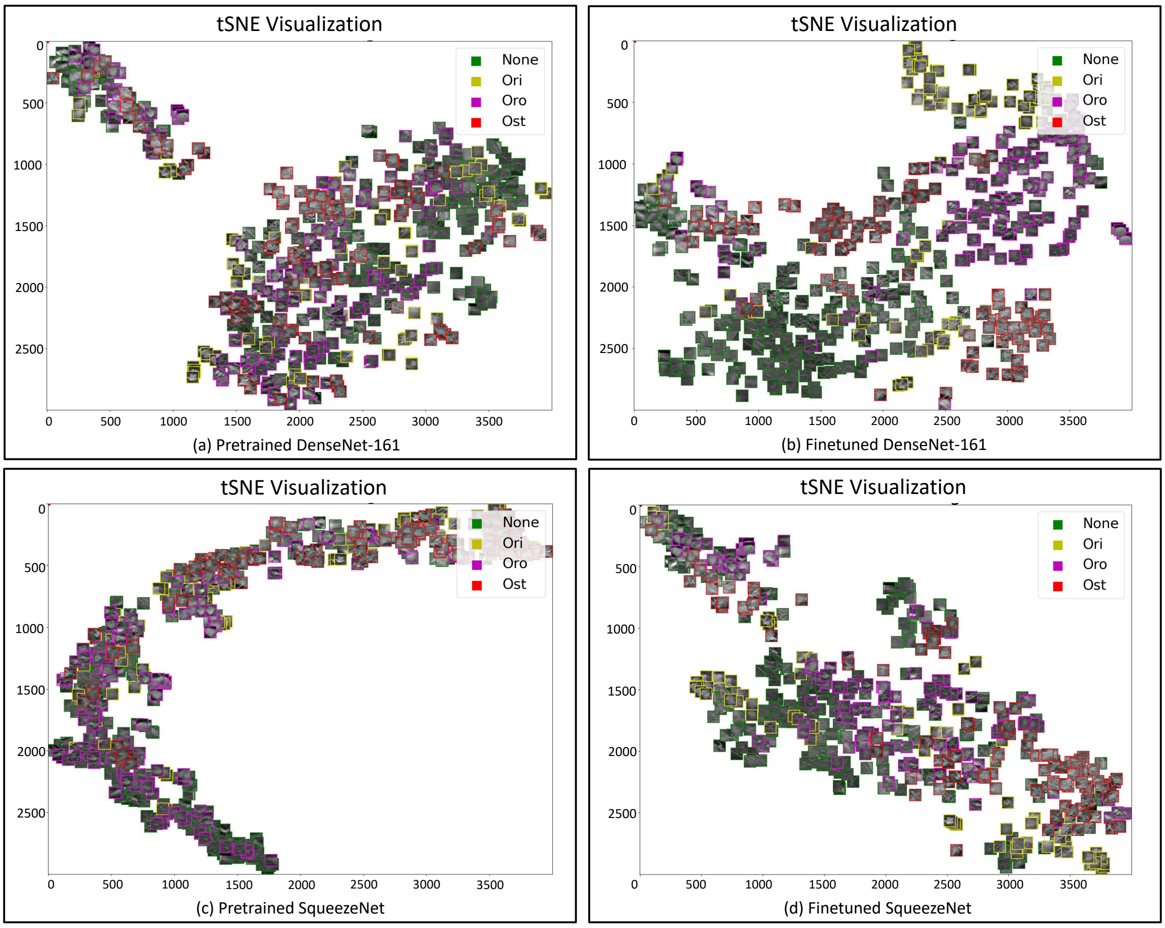
7.4.1 t-SNE and UMAP

The extracted features from both pretrained and fine-tuned networks are visualized in order to understand what patterns emerge in low-dimensional spaces after having employed nonlinear dimensionality reduction techniques such as t-SNE and UMAP.

In Figure 7.3, the t-SNE embedding plots for both pretrained and fine-tuned DenseNet-161 and SqueezeNet architectures are pictorially represented. Similarly, Figure 7.4 presents the UMAP embedding plots for both pretrained and fine-tuned DenseNet-161 and SqueezeNet models. In the pretrained version, no clear patterns arise from both embedding plots, showing that features learned from ImageNet dataset are not necessarily well discriminative for radiological image applications.

Nonetheless, after 50 epochs of fine-tuning on the designated train set, the clusters appear more distinctive. In fact, with trained CNN features, both UMAP and t-SNE allow to visualize different clusters for all four considered classes: *None*, *Ori*, *Oro*, and *Ost*.

As described in Section 7.3, the distance between the clusters cannot be interpreted by using the t-SNE visualizations. For instance, it cannot be inferred from Figure 7.3 that clusters are dissimilar to each other when one cluster is closer to the other. However, it can be stated that points closer to each other are more similar objects than the points at farther



(a) Pretrained DenseNet-161

(b) Finetuned DenseNet-161

(c) Pretrained SqueezeNet

(d) Finetuned SqueezeNet

Fig. 7.3 The t-SNE embedding plots of the features extracted from pretrained (**a**,**c**) and fine-tuned (**b**,**d**) DenseNet-161 and SqueezeNet, respectively, on the validation set of 1st fold. It is clearly visible that the fully TL paradigm does not allow a clear clustering of the features in low-dimensionality space, whereas the finetuned model is able to discover more discriminative features with respect to its pretrained-only version.

ends, whereas Figure 7.4, thanks to the local and global feature representation capability of the UMAP, clearly plots the points that can be interpreted as distinguishing clusters and the position of the points.

7.4.2 Class Activation Mapping

The visual explanation of all eight fine-tuned networks is pictorially depicted in Figure 7.5, considering the Grad-CAM as reference method. In the figure, two sample images for every class are depicted, and the corresponding saliency maps are reported for every network. This figure considers only images for which every network makes the correct prediction,

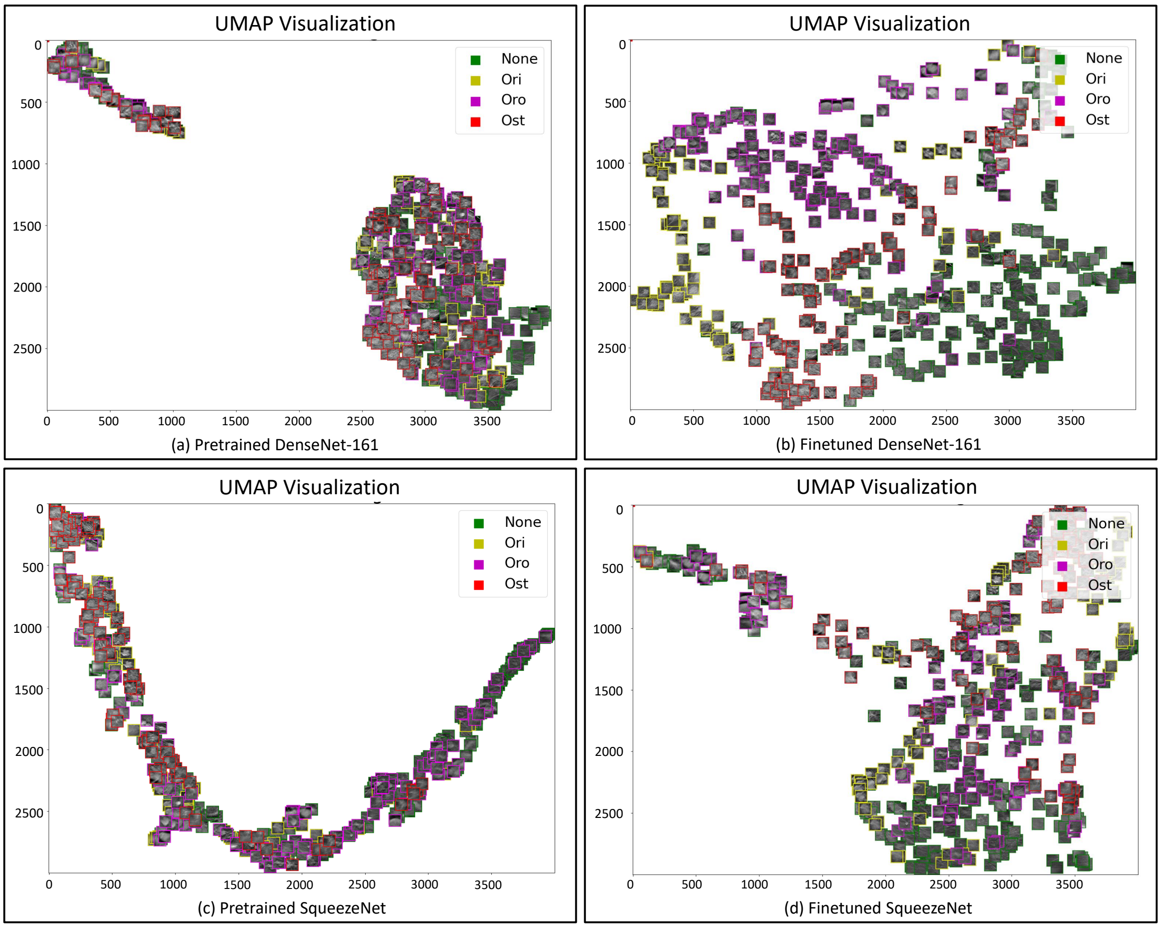


Fig. 7.4 The UMAP embedding plots of the features extracted from the pretrained (**a**,**c**) and fine-tuned (**b**,**d**) DenseNet-161 and SqueezeNet, respectively, on the validation set of 1st fold. It is distinctly visible that the fully TL paradigm does not allow a clear clustering of the features in low-dimensionality space, whereas the finetuned model is able to discover more discriminative features with respect to its pretrained-only version.

in order to visualize the link between the highlighting of the lesion area and the network performance. The saliency maps of the approximate features are generated considering the ground truth/predicted class view.

Interestingly, the CNN architectures that find troubles in correctly identifying the lesion areas also appear to have worse performance in the classification task. For instance, SqueezeNet, which is the worst-performing network in terms of AUC, and VGG-16, which also appears to have a trade-off between AUC and the number of parameters, as shown in Figure 6.6, fail to spotlight the relevant lesion area. Here, the trade-off refers to the fact that the increasing number of parameters seldom yields increased AUC. In contrast, DenseNet-161, DenseNet-121, and ResNet-50 correctly highlight the lesion on the images.

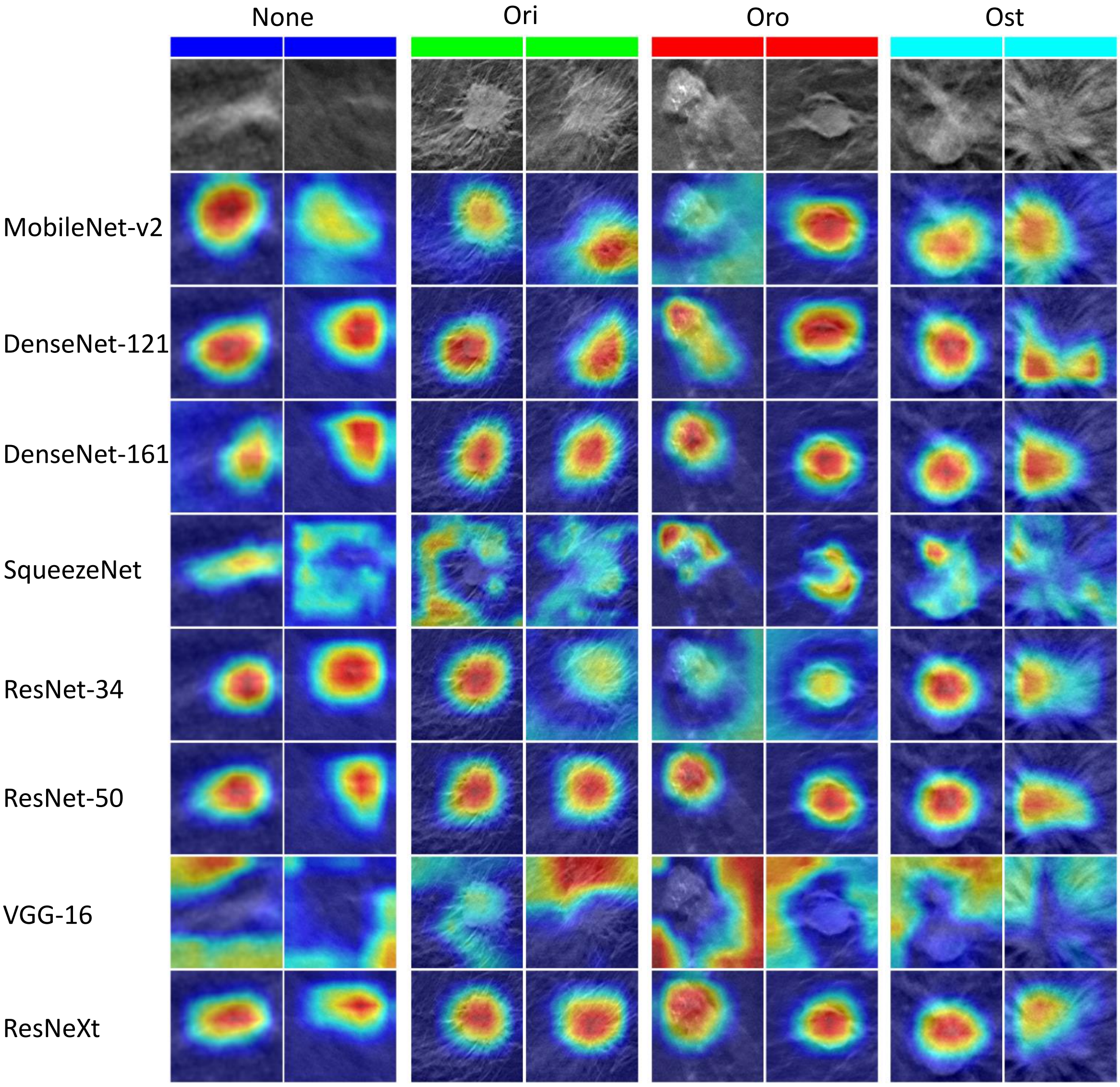


Fig. 7.5 The visualization of the Grad-CAM method with the eight different CNN architectures considered throughout the study. To illustrate the better view, two examples for each class are portrayed, and the ground truth class label is provided above the set of each image. As the jet color scheme is employed for depicting saliency zones, the red color represents the higher intensity, i.e., pixels on which the network is focusing more for performing the classification, whereas the tendency towards the blue color represents the lower intensity of focus. The header bar is used to distinguish among several classes and is colored uniquely. The similar color of header for two images represents the samples chosen from the same class.

Thus, this XAI-based CAD system unveils the potential applicability of the reliable and suspicious candidates to adopt in the CAD systems.

7.4.3 Local Interpretable Model-Agnostic Explanations

It is worth mentioning that the visual results of the Grad-CAM and LIME must not be confused. Unlike the Grad-CAM method, which emphasizes the lesion area with the intensity of the color closer to the center, the LIME method works differently by providing the top contributing s that resulted in the classification of the image into any given class. However, in both cases, the images were generated by observing the ground truth/predicted class view.

The s perturbations performed by the LIME are shown in Figure 7.6. The observations experienced with respect to the performance of the LIME technique are similar to the Grad-CAM method. The figure reports the exact images that were compared in Figure 7.5 for the Grad-CAM method, to create a robust and clear comparison. The class considered for performing the LIME perturbations is the ground truth class, which, in this case, corresponds also to the prediction of all the CNNs. The regions which are positively correlated with the decision made by the CNN are highlighted in green, whereas those negatively correlated are colored red.

However, it has to be noted that reasoning in terms of superpixels can result in explanations which are visually less clear to understand than those of their CAM-based counterpart. Comparing the Figures 7.5 and 7.6, one can see that some superpixels which are correlated to the prediction according to the LIME method are not considered relevant in the corresponding Grad-CAM activation maps. Therefore, the study suggests to consider both methods when trying to devise an explanation for a CAD system, in a way that complementary information can be extracted from both sources to obtain a broader view of how the model is working.

7.5 Discussion

This study proposes a novel, visually explainable DL-driven multiclass shape-based breast cancer classification framework for tomosynthesis lesion images, as presented in the Chapter 6. For the task of morphological classification, eight DL models are employed on tomosynthesis breast images. However, the blackbox nature of DL hides the decision making mechanism, which hinder the incorporation of DL in medical domain. Therefore, to build the necessary trust of physicians and health experts, the explainability of the DL driven CAD systems is essential.

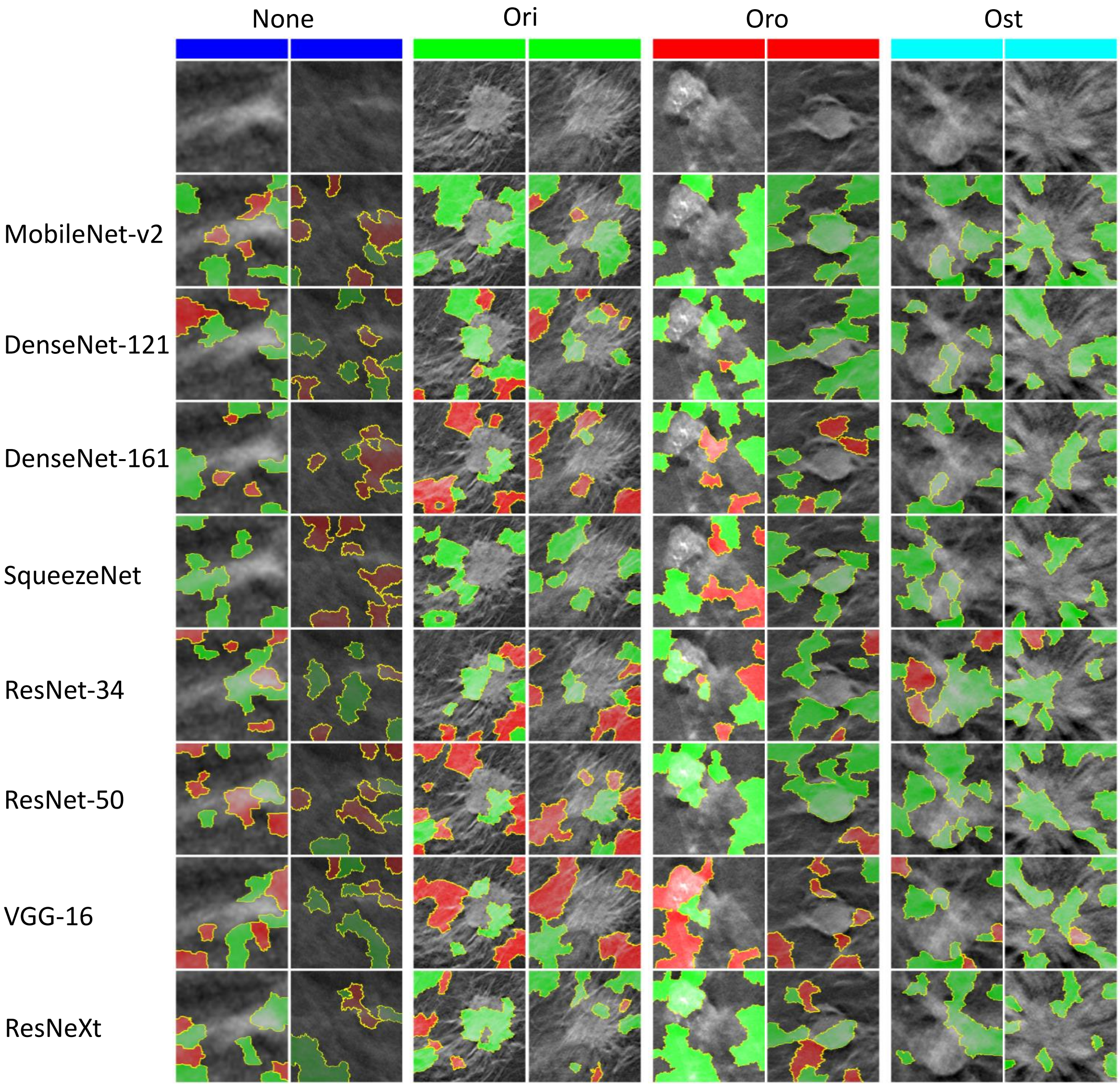


Fig. 7.6 The visualization of LIME superpixels positive and negative regions with the eight different CNN architectures considered throughout this study. To illustrate the better view, two examples for each class are portrayed and the ground truth class label is provided above the set of each image. The red color highlights the negatively contributing superpixels, whereas the green represents otherwise. The header bar is used to distinguish among several classes and is colored uniquely. The similar color of header for two images represents the samples chosen from the same class.

In this context, two families of XAI methods, i.e., perceptive interpretability and mathematical interpretability, are incorporated to explain the results acquired during the validation study of breast classification discussed in the Chapter 6 in order to create the trust among the clinicians and the AI.

The perceptive interpretability models are responsible for visually explaining the top contributing features towards the classification, whereas the mathematical interpretability methods portray feature clustering capabilities of the DL architectures.

The CAD system developed in this study is able to improve the diagnostic and prognostic performances. The successful implementation also enhances the trustworthiness among the clinical field and the high-accuracy-yielding DL architectures. The sections below comparatively discuss the interpretation of the DL models using XAI techniques.

7.5.1 Explainable AI in Breast Cancer Classification

Concerning the mathematical explanation, as emerged from the visualization of the feature embeddings, one can discern that both t-SNE and UMAP are able to extract meaningful relationship in the low-dimensionality spaces when the features are representative of the underlying patterns in the sample images. In Figures 7.3 and 7.4, four clusters are clearly visible for the DenseNet-161 architecture. On the contrary, when the model is less accurate, as in the case of fast and light SqueezeNet (in terms of number of parameters), the cluster formation behaves differently, with UMAP resulting in more compact representations. As a general suggestion, therefore, the study recommends to use these mathematical XAI techniques to visualize if considered features for a problem under consideration are relevant.

With respect to the perceptive XAI techniques, the performance results of the CNN models are aligned with the complementary information that can be extracted from Grad-CAM and LIME methods. While the first allows to detect which regions have a gradient that is deemed relevant for performing the prediction, the second permits to understand, for each superpixel, if it is positively or negatively correlated to the prediction. Moreover, the LIME method has an adjustable parameter for deciding the number of top contributing features to show over the original image. Since the intensity values from the saliency maps of Grad-CAM are already available, every positively correlated region is marked in green color and every negatively correlated region is highlighted with red color, so that the mixed information obtained can be exploited to obtain an intuitive understanding of which regions are more important (higher intensity values in CAM maps), and which are positively or negatively correlated to the final outcome (green and red, respectively).

Interestingly, the CNN architectures that find trouble in correctly identifying the lesion areas also appear to have a lower AUC. Thus, on a general scale, the higher AUC can be explained by using XAI methods. For instance, the SqueezeNet, which is the worst-performing network in terms of AUC and validation loss, and VGG-16, which has a trade-off between the AUC and the number of parameters, as shown in Figure 6.6, fail to spotlight the relevant lesions as illustrated in Figure 7.5. In contrast, DenseNet-161, DenseNet-121, ResNeXt, and ResNet-50, which feature higher AUC values, correctly highlight the lesion when tested with the Grad-CAM method.

Moreover, as the loss trends and the AUC tables show, none of the CNNs yielded 100% performance, which means the misclassified examples are also present. These samples of the misclassified images are also presented to the XAI methods in order to dive into the features that resulted in misclassification. The reason behind misclassification of one type of cancerous image to another type might be related to the homogeneity of the shapes of a few examples with other classes. Figure 7.7 illustrates the results of both Grad-CAM and LIME methods regarding examples of misclassified images.

The labels provided above the samples represent the ground truth, whereas the labels provided under the saliency maps are the predictions made by the CNNs. This figure proves that XAI could also help the physician understand why the AI is failing.

For instance, the *None* image in the Figure 7.7 contains a mesh that is not lesion according to the expert radiologists. However, it fools the CNN to misclassify the image as *Ori*. Both the CAM and LIME methods highlighted the regions that carry analogous properties, thus explaining the cause of the misclassification. It is worth noting that a similar discussion emerges from the other examples provided in Figure 7.7.

In order to understand how the results of two XAI perceptive methods vary according to the different target classes, Figure 7.8 reports the explanation results of both methods considering eight different correctly classified images. It is worth noting that the CAM results did not differ among the four XAI target classes. Interestingly, the results are different when LIME is analyzed. For instance, when considering *None* as target class and visualizing its explanation outcome on an *Ori* class image, the LIME-highlighted lesion area has a region that contributes negatively towards the classification of the chosen target class. In the same image, the LIME explanation with the *Ori* target class highlighted the lesion region as green (positively correlated) since the image belongs to the *Ori* class. This kind of comment could also be easily applied to other images of Figure 7.8, thus confirming the difference and the utility of more than one perceptive XAI method.

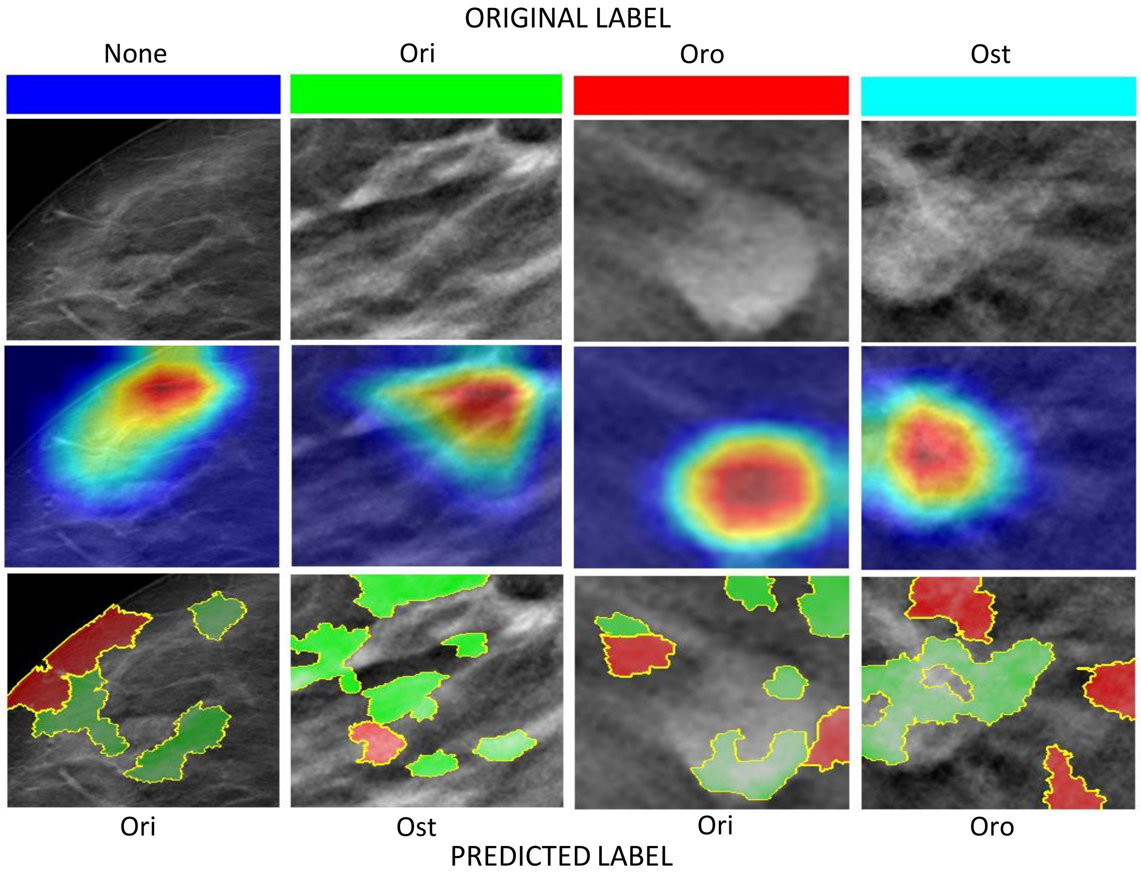


Fig. 7.7 The examples of the misclassified samples due to the relevancy of one type of shape to other type of shape for all four classes. The labels provided above the samples represent ground truth, whereas the labels provided under the saliency maps are the predictions made by CNNs.

Finally, from several discussed dimensions, the presented study proves the applicability of XAI methods and the black-box nature of the DL models is successfully unveiled to build the trust of radiologists to emerge towards the reliable CAD systems for the diagnostic tasks.

7.6 Limitations and Future Directions

The study unleashes the hidden classification mechanism of DL techniques by integrating numerous XAI techniques. The Grad-CAM methods produce a coarse localization map. In the experimental outcome section, it can clearly be observed that at a certain point the XAI methods explain the DL methods' results with slightly different regions. This is because

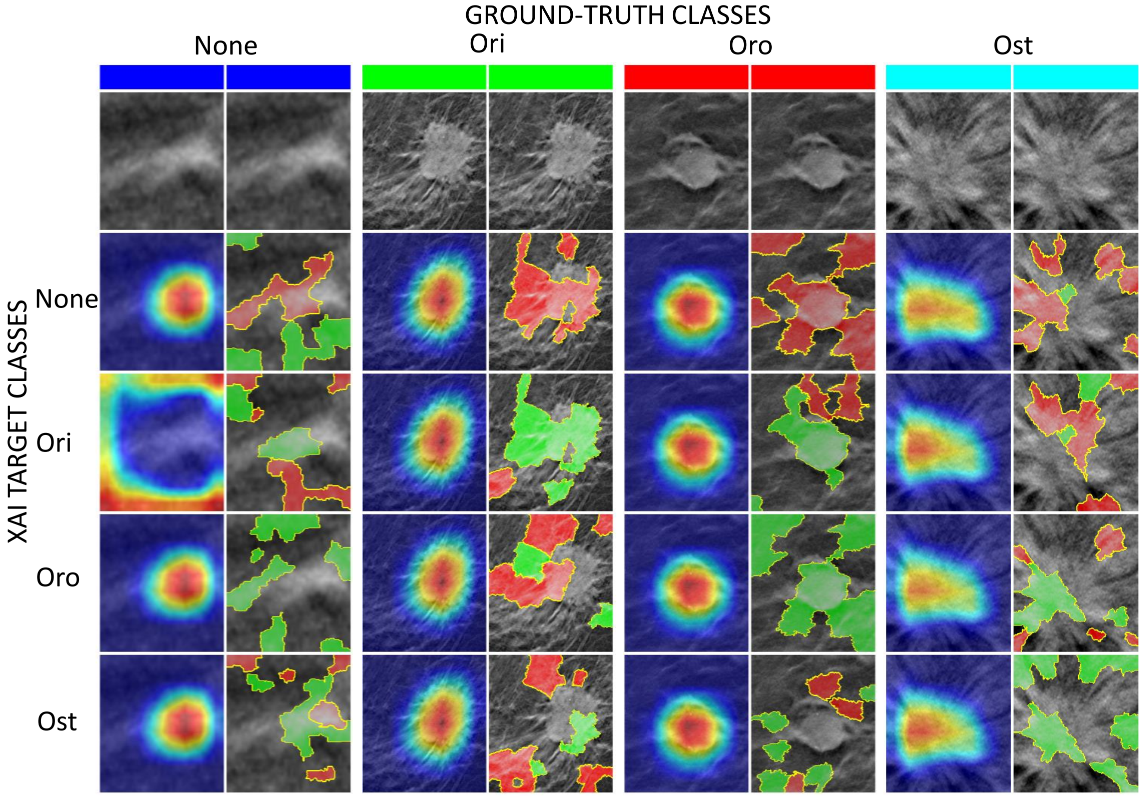


Fig. 7.8 Grad-CAM and LIME comparison. One sample image is used for every class; then, the results of the XAI perceptive method are shown, considering each of the four possible target classes.

of the model overfitting. A robust investigation may demonstrate productive conclusions. The CAM method only focuses on a general region of the image instead of focusing on minute peculiarities, such as that LIME technique generates the perturbations and highlights the top features. Similarly, the SHapley Additive exPlanations (SHAP) model quantifies the exact amount of contribution made by a particular region, and can be added in future studies.

More importantly, there exists no particular method to quantitatively and qualitatively evaluate the outcomes of XAI methods. Merely visualising the heatmaps and portraying the top contributing features may not make a DL model completely interpretable in medical diagnosis and treatment context. Therefore, investigating quantitative and qualitative measures to evaluate the results of XAI techniques is the precise future target.

7.7 Summary

In this work, a robust visually and mathematically explainable DL framework for multiclass shape classification of tomosynthesis breast lesion using eight pretrained CNN models using an in-house dataset is proposed. Due to small-scale data availability, the data augmentation was incorporated. The best fine-tuned model achieved mean AUC values of 98.2% and 96.3% with and without considering the data augmentation, respectively.

Furthermore, considering the hypersensitive clinical realm, two families of XAI methods, i.e., perceptive interpretability and mathematical interpretability, were incorporated to visually explain the CNN models' classification performance. The former interpretability method includes Grad-CAM and LIME, which are responsible for visually explaining the experimental outcomes in terms of feature-level contribution towards classification, whereas the latter method comprises t-SNE and UMAP techniques that portray feature clustering capabilities of the DL architectures. The performances of all models were aligned with the visual and mathematical interpretations, hence developing the necessary trust between the healthcare industry and the DL architectures. The results proved the usability of XAI to understand the mechanism of employed AI models, also in the cases of failures.

In future, the aim is to further enhance the interpretability of the CNN models by calculating the single feature-level weightage towards classification. The other plan is to investigate the performance of the proposed framework on unforeseen datasets and to integrate the novel DL models. However, more importantly, investigating quantitative and qualitative measures to evaluate the results of XAI techniques is the precise future target.

Chapter 8

Evaluation of Explainable Artificial Intelligence

8.1 Visualisation is not Explanation

The DL has started romance with almost all fields of human life and beyond. None of these fields know how DL works and many seldom bother because of the perks DL offers. However, the medical domain has a suspicion that DL may cheat in this romantic partnership. Therefore, before indulging into serious relationship, healthcare demands some sort of guarantee and assurity in terms of interpretability, transparency, explanation, and evaluation of the DL methods. Eventually, the XAI, a good mediator, has offered some degree of transparency, mainly by visualisation, to ensure trustworthiness. The healthcare industry, sensitive and gorgeous darling because of having the direct correlation with human life, questions the authenticity of XAI, and demands a qualitative measure to evaluate the explainable methods. Therefore, in the greater interest of XAI led DL relation with medical domain, this part of the study investigates the evaluation methods and measuring metrics to quantitatively and qualitatively evaluate the XAI methods in medical imaging domain. The retrospective survey focuses beyond the visual explanations and saliency maps and further investigates the quantitative and qualitative methods to judge whether the explanation itself is worth trusting. Most of the DL based studies in medical images compute the heatmaps and claim the models to be interpretable. In the author opinion, the visualisation is not explanation. Very few studies go beyond and adopt methods like intersection over union, which is most commonly used performance evaluation metric. This retrospective study found that there exists no single generalised quantitative method to evaluate the XAI in medical imaging domain. Therefore, the author expresses the dire need for a generalised method to quantitatively and qualitatively interpret the explainable methods within XAI domain for medical images and look forward to the scientific community.

8.2 XAI Evaluation Background

In literature, several methods have been proposed for the evaluation and quantification of XAI methods, however, there is no one compact and generalised method for quantitatively evaluating the XAI outcomes on different types of medical images. A common practice has been seeking help from the clinicians to evaluate the explanations generated by XAI methods, however, this method is prone to errors, time consuming, labour intensive, and experience demanding. Nevertheless, the visualisation of top contributing features, spotlighting the important regions, and computing numerous scores of contribution towards decision have long been discussed, alongside what is required is a quantitative and qualitative method to measure the effectiveness of an explanation.

Additionally, as the definition of XAI states, the purpose of XAI is to make the DL decisions understandable to human. A number of studies merely relying upon saliency maps and feature contribution values lack the actual definition. The debate to make the XAI decision understandable to an expert or a common human can be considered progressive, however, the requirement to explain remains intact in either case.

Moreover, unlike other domains, the medical imaging domain can not rely on common sense, mutual understanding, and argumentation about the XAI methods, therefore, very limited number of studies have been found interested in evaluating the effectiveness of the XAI techniques. Numerous authors provide different kinds of evaluation including visual, textual, example based, and few more.

8.3 Medical XAI: A Quick Look

This section further elaborates the categories of XAI methods applied on medical image diagnosis and analysis using the DL. The literature study suggests few major groups including post-hoc XAI vs model based XAI; model specific XAI vs model agnostic XAI; and global explanation vs local explanation, as explained hereunder. The in-depth picture and exhaustive review of the XAI methods and applications do not fall in the scope of this study, and can be found in the literature [428, 429].

8.3.1 Post-hoc XAI vs Model-based XAI

The prominent difference between post-hoc and model-based is the approach both models follow. The former is applied to a trained DL model and the insights (i.e. feature learning, features importance, other model behaviour, etc.) of the DL model are explained mainly using the saliency maps and other visualisation techniques [430]. Several post-hoc models also work by perturbing the input to understand the significance of a particular feature towards output, and the degree of contribution towards classification.

The latter aims at making the model more interpretable. The traditional machine learning techniques e.g. decision tree, support vector machine, and regression fit onto the model based explanation, where the process between input and the output remains linear and explainable [430]. These models are also commonly known as intrinsic models due to their white-box nature (i.e. results are explainable and understandable to human). As stated in the introduction section, this study targets to explain the blackbox behaviour of DL models, therefore, the model-based XAI techniques are not given priority.

8.3.2 Model Specific vs Model Agnostic XAI

This interpretability definition of model specific versus model agnostic comes from the limits on the selection of models they can be applied to. Model specific approach can only to be applied to a particular set of architectures that allow interpretability models to access and alter (if required) the internal working mechanism of network. This access to the internal information is not easily available on vast majority of DL models, therefore, making the model specific explanation less desirable choice in practice. Additionally, according to Adadi et al. [431], all model based explanation are by default also model specific, however, a model specific explanation may not necessarily be the model-based explanation. This claim can be justified by the applicability of certain class of post-hoc methods, i.e. saliency mapping models, that are applicable to only particular class of CNN.

Conversely, bearing no restriction, the model agnostic methods do not require the selection of a certain network from a pool of neural networks, but allow a open range of choice for network selection [432]. The model agnostic explanations are more concerned with the input to the output and hence, perturb the input to acquire the information about the importance of a particular region and contribution to the decision.

8.3.3 Global Explanation vs Local Explanation

This subsection defines the scope of an explanation provided by an XAI methods over the network. The global explanation, also referred as dataset level explanation, yields insights into the learned behaviour of the whole network [433]. A global explanation reveals what algorithm has learned in terms of features, feature importance, and upto what extent. An example consideration can be, the same set of features extracted against all the images in a particular dataset. These features are advocate of the decision made by the architecture.

On the other hand, local explanation describes the output of the model against a single instance of the input. Here, the behaviour of the model is examined against one example that has been classified into a certain class [434]. An example of the local explanation model can be considered as heatmap generation on the localisation of a breast tumor on a single tomography image.

8.4 Explainability Methods in Medical Imaging

There exist several explanation methods, which in reality are visualisation methods to merely highlight and showcase the important regions. These methods include various CAM based techniques, a perturbation method called LIME, a game theory based approach named SHAP, a propagation method titled LPR, and many more. Few of these methods and salient distinctions are briefly described in the below subsections.

8.4.1 Grad-CAM

The Grad-CAM, a post-hoc explanation method, is a generalised variant of CAM method proposed by Selvaraju et al. [425]. Unlike CAM method, that requires global average pooling by replacing fully connected layers, the Grad-CAM can be applied to anyof the CNN architectures for heatmap generation. The same authors also proposed an extension of Grad-CAM called Guided Grad-CAM that works on element wise multiplication.

One of the most important points to note is that attention maps generate positive and negative values and the keeping both types of values is crucial to understand the contribution for decision prediction. The CAM methods imitate the behavior of ReLU activations and ignore the negative values coming from the attention maps, whereas, the Gradient based techniques focus on calculating the absolute value.

8.4.2 LIME

Another famous model agnostic explanation method called LIME is proposed by Ribeiro et al. [435]. The LIME model works by generating several local explanations of the complex model, perturbing the input, making several different modifications in the input, and approximating the complex model into a linear model. The usage of LIME method in medical images remains simple and quite effective, however, similar to other explanation methods, LIME too can be fooled and it also suffers from adversarial attacks [436].

8.4.3 LPR

The LRP, introduced by Bach et al. in 2015 [437], highlights the important regions in the voxels by generates the heatmaps and computes the computes the classification score that ranges between 0 and 1. This classification score, sometimes referred as relevance score, propagates back to the network and directly spotlights the positive value for the classification decision.

8.4.4 SHAP

Another model agnostic explanation approach is proposed by Lundberg et al. [438]. The SHAP technique is based on the Shapley values of game theory approach. The SHAP generates the values that compute the contribution of the desired features (usually top contributing features). These features are evaluated individually. However, similar to LIME, the SHAP is also prone to the adversarial attacking problem [436].

8.5 Evaluation Measures and Metrics for XAI

In literature, several different explanation types are generated i.e. visual, textual, example-based, etc., therefore, the the measures and metrics vary depending upon the explanation type provided. Moreover, within one type of explanation e.g. visual explanation, numerous explanation methods exist, thus the metrics to quantify the explanation may change from modality to modality and model to model. Additionally, it is worth mentioning that context of the application matters. For instance, an XAI method for breast cancer classification network on 3D images must not be evaluated on the same grounds as an XAI method for report generation for whole slide images. In this study of evaluation of XAI for medical images, the most common interpretations are visual and textual explanation. The subsections below further shed some light on the topic.

8.5.1 What is Quality/Enough Explanation?

Explaining the explanations has been well searched topic in recent years, however, one of the most proficient and prudent questions is to define what is good explanation. What defines/declares and makes an explanation good is another relevant and interesting question to raise. Several studies have been conducted to answer these question and to define the criteria for goodness of an explanation [439, 440].

Additionally, the explanation of explainable methods is also context dependent that arises another question, i.e. an explanation must be understandable and interpretable but to whom? To general public? To experts? Or to whom? An interpretation of the cancer classification model on breast images is only understandable to physicians and relevant experts. These queries are addressed in an interesting study by Tomsett et al. [441]. Therefore, all these questions open new horizons and direct to the context dependent applications. The evaluation of the XAI methods depends upon the end user of the application and the sufficiency of the quality of explanation depends upon the application area, explanation purpose, and the targeted audience. The coming sections define some metrics in the context of visual and textual explanation of XAI methods applied on medical images.

8.5.2 Measures for Visual Explanation

In light of the aforementioned concerns, the most commonly applied quantitative measure to objectively evaluate the heatmaps and attentions generated by XAI models for visual explanation has been IoU [33]. The saliency maps highlight the important regions on the images, these maps are compared to the ground-truth images, and the intersecting regions is measured to quantify the performance.

Another similar method to assess the performance of saliency maps is to compute the area of the perturbing curve with respect to the first most relevant perturbation is proposed by Samek et al. [442]. The strategy is somehow similar to the LIME method in a way that the method perturbs the different regions over repetitive iterations and computes the sensitivity, importance, and relevance of the regions towards the decision.

A delete and insert strategy to evaluate the performance of XAI models in an effort to explain the blackbox neural networks is proposed by Petsiuk et al. [443]. The method computes the probability of each pixel towards the class using the deletion and insertion

Table 8.1 The recent survey articles and their highlights

Objective	Year	Evaluation	XAI			Explanation Type		Ref
			Visual	Textual	Others	Post-hoc	Ante-hoc	
Explaining XAI in medical images	2020	✗	✓	✗	✗	✓	✓	[34]
XAI for healthcare	2020	✗	✓	✓	✓	✓	✓	[446]
DL based XAI for image analysis	2020	✗	✗	✗	✓	✓	✓	[447]
Medical image interpretation	2021	✗	✗	✓	✗	✗	✓	[448]
XAI in image cancer detection	2021	✗	✓	✗	✗	✓	✓	[33]
Interpretability of neural networks	2021	✗	✓	✗	✗	✓	✓	[449]
DL based XAI for image analysis	2022	✗	✓	✓	✓	✓	✓	[450]
XAI in medical image diagnosis	2022	✗	✓	✓	✓	✓	✓	[451]
Evaluating XAI for Xray images	2022	✓	✓	✗	✗	✓	✓	[452]

metric and discards the least probable pixels in terms of information yielding regions. The insertion and deletion metrics were used to calculate the increasing and decreasing probability of a pixel towards explanation, respectively. However, the strategy has not been quite successful because of single measure to calculate the model performance degradation. The model can also reveal poor performance in the presence of artefacts, bias, and imbalance in the data.

Following the similar drawback, a remove and retrain strategy named ROAR (RemOve And Retain) has been proposed by Hooker et al. [444], in which the features on an image were randomly drooped and the accuracy of the model was rechecked. The features that brought significant decline in model performance were retained and other discarded and so on. The model required several iterations of retraining and reevaluation.

Similarly, Eitel et al. [445] presented positive and negative relevance scores of the lesion areas towards the classification and decision making. A higher value of the relevance score advocates the importance of the regions in classification.

In spite of the availability of several evaluation measuring metrics, the requirement for a well established and generic protocol to assess performance of the saliency maps is intact, which invites the scientific research community to develop one.

8.5.3 Measures for Textual Explanation

The textual explanation involves the interpretation of caption generation and report generation on medical image analysis and diagnosis. The similar domain also falls under the category of natural language processing, therefore, the most commonly applied XAI evaluation metrics comes from the language processing domain. These metrics are sometimes used along with saliency maps to further strengthen the explanation claim.

The most common performance measuring metric is Bilingual Evaluation Understudy (BLEU) that computes the matching of the generated text with the ground-truth [453]. The BLEU score is computed between the range 0 and 1, where 1 means a 100% match with the ground-truth. The similar score is calculated over a range of 'N' iterations of BLEU score. In this study, the author found that BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are most most commonly computed values with 4 iterations.

Another well-known quality assessment measure for the text explanation is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [454]. The ROUGE is pretty old method, in fact before the modern advent of XAI methods, however, it still remains active due to its applicability. It compares the generated text with the ground-truth text (sometimes referred as reference text) in terms of length of common string and computes recall and precision. The longest the string in the ground-truth and the generated text, the better the explanation is.

Unlike other approaches, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) works by assigning a particular weight to the position of text or a term in the document [455]. Lastly, a very well-known and frequently employed technique in many other domains, including the compression algorithms and transcription, is Term Frequency Inverse Document Frequency (TF-IDF). A TF-IDF based Consensus-Based Image Description Evaluation (CIDEr) approach is proposed by Agarwal et al. [456]. It works on a very simple and intuitive mechanism that the terms that appear quite frequently (i.e. is, am, also, are, etc.) may not be as important as other terms that appear regular but less frequent.

8.6 XAI Evaluation Literature

This section presents the studies that incorporate different local methods to explain and evaluate the outcomes of XAI methods in medical imaging tasks. The summarised results of the studeis are also presented in the Table 8.2.

8.6.1 Visual Explanation Evaluation

A weakly supervised multiple brain lesion detection method on MRI images using UNet architecture is presented by Dobust et al. [457]. Several CAM based visualisation methods are adopted to attain the attention maps. The 2D slices are extracted from the attention maps to compare with the annotated images. The Hungarian algorithm is used to match the lesion with the annotation. Considering the intrinsic nature of the Hungarian algorithm, that yields the confidence scores, the area under the ROC curves is presented as a measure to evaluate the results obtained by XAI methods. The authors presented numerous experiments with varying structures of UNet model, therefore, the average ROC reached 72.0 ± 13.3.

Eitel et al. [445] presented an explainable network for the diagnosis of multiple sclerosis on MRI images using 3D-CNN and the well-known LRP for the explanation. Authors proposed and calculated relevance score (both positive and negative) to the lesion area to quantify the explanation generated by XAI method. The mean and the standard deviation remain in the range of $-1.05e-60.0013$ for negative relevance sum in comparison of 3.07e-06 ± 0.0014 for a sum of positive relevance. The sum of the relevance of the area for the multiple sclerosis patients was 9.71%, of which 5.15% was attributed to lesion area.

The authors in [458] proposed a 3D CNN based framework for the detection of coronary artery atherosclerosis in CT images. The output of the framework, i.e. discriminatory features, are visualised using Grad-CAM based saliency map at the final convolutional layer. To explain and quantitatively evaluate the highlights of XAI based Grad-CAM method, the authors computed the pixel level overlap of ground-truth image with the model prediction and achieved the results of Dice 0.58, accuracy 0.63, and sensitivity 0.77. Further information on how the definition of these metrics is slightly altered for the computation purposes is provided in the article [458].

An DNN based classification pipeline on Xray images to identify the known as well as unknown diseases in COVID-19 patients is presented by Tang et al. [459]. The CAM and a variant named DisCAM explanation methods are applied on two datasets (i.e. skin lesion and chest Xray images). Similar to the aforementioned methods, the evaluation of the CAM method is performed by comparing the highlighted regions of CAM method to the ground-truth images. The article claimed that the CAM method does not necessarily spot the important regions for the unknown disease classification. This fact also questions the implementation of the DNN model in the given context.

An attention based CNN workflow for the glaucoma detection along with glaucoma database are developed by Li et al. [460]. The attention maps are computed not only to spot the salient regions but also to improve the detection of glaucoma. The attention maps and the CAM methods are applied and the Pearson correlation coefficients are computed on the ground-truth images to testify the performance of XAI methods. The correlation coefficient reached the value of 0.581 and a variance of 0.028. Apart from the transparent model, the detection accuracy and AUC values were also pretty impressive during the experimental phase.

Similarly, another attention maps and DL driven stenosis localization and classification framework on coronary angiography images is proposed by Cong et al. [461]. For the sake of evaluating the employed stenosis activation maps explanation method, the mean squared error and the localization sensitivity with respect to the generated bounding boxes are computed as measuring metrics. The sensitivity score reached 0.72, whereas, the mean squared error remained 69.6 for the stenosis positioning.

Another LPR for DNN based so-called explainable approach for the classification of Alzheimer's disease on MRI data is introduced by the [462]. Similar to Grad-CAM visualisation technique, the LPR generates the heatmaps over the salient regions/features that positively contribute towards the classification and localisation decision. The regional overlap between the ground-truth and the prediction and the sum of importance of the Alzheimer's disease area were considered two measuring metrics to quantitatively evaluate the heatmaps indicating the relevance of image to disease. The impressive results reached over 90% accuracy in accessing and visualising the outcomes.

korbar et al. [463] presented an explainability oriented ResNet architecture to indicate the salient features and regions that led to classification task in the whole slide images of the colon polyps. The Grad-CAM and Guided Grad-CAM visualisation techniques are used to generate heatmaps, and the pixel area to be inside the RoI, and IoU are considered the evaluation metrics for XAI. The achieved results of IoU for the Grad-CAM, Guided Grad-CAM and Guided Grad-CAM with boxes reached 0.24, 0.47, 0.55, respectively.

A model agnostic multi-scale segmentation network based on CNN techniques to visually explain the results is explained in the study of Seo et al. [464]. The authors proposed a generic pipeline for evaluating the XAI results, especially CAM methods, on several different types of images. The maximum IoU and the mean IoU are computed for quantitative evaluation of the explainable method and the heat maps generated. The maximum IoU remained 55, whereas, the mean IoU reached 42.9 over numerous thresholding values.

A UNet architecture to segment the craniocerebral regions of transventricular and transcerebellar fetal brain planes on ultrasound images is studied by Xie et al. [465]. The famous Grad-CAM visualisation method is employed to spotlight the important regions containing the lesions. The model was run on the expert annotated dataset and the bounding boxes are predicted for only abnormal images by fitting the network's learned RoI over the ultrasounds. The IoU on the predicted bounding boxes and the ground-truth images is calculated to statistically evaluate the XAI method. The standard deviation and the mean IoU on lesion localisation are 0.497 and 0.126, respectively.

An investigation of different types of biases towards the binary classification to discriminate and explain the tumour tissue is studied by Hagele et al. [466]. The incorporation of public dataset of haematoxylin-eosin-stained images and the LRP method to provide pixel level explanation are what made the system explainable. However, as the authors clearly stated, visualisation merely seldom brings any explanation, therefore, the ROC curves are computed to evaluate the explanations. Since the LRP yields pixel level explanation, the relevance value for each cell is computed, and the resultant AUC reached 0.76.

Each pixel carries some information towards the classification and detection, for which, several traditional methods have been proposed in literature. A novel visual attribution approach, unlike existing ML and DL approaches, uses Wasserstein Generative Adversarial Networks (WGAN) to study the interpretability on mild cognitive impairment and Alzheimer's disease on publicly available images [467]. The WGAN based technique features the capability to detect and label the specific area on the image to the relevant class. The authors computed normalised cross correlation as an evaluation metric to the XAI method. The visual attribution WGAN achieves better results than competitors CAM, Integrated Gradients, and Guided Backprop with the mean normalised cross correlation and standard deviation of 0.07 over the ground-truth and the predicted maps.

Lin et al. proposed ResNet-50 based deep CNN to provide an explainable and interpretable model for multiple image types [468]. The authors employed several different XAI methods including LIME, SHAP, Gradients, and GSInquire to evaluate the performance of DL model and the visualisations. Two statistical measures, namely Impact Score and Impact Coverage, are introduced to quantitatively evaluate the outcome (saliency maps and/or region highlighting) of XAI models. The acquired results for Impact Score and Impact Coverage against LIME: 38.05% 35.12%, SHAP: 44.15% 40.24%, Gradients: 51.22% 47.80%, and GSInquire: 76.10% 50.73%, respectively, remained satisfactory.

8.6.2 Textual Explanation Evaluation

The explanation of DL models on medical images is not solely provided with visual techniques like heatmaps and attentions. Several textual representation and example oriented explanation methods are also described in the literature. Spinks at al. presented a neural network for visual and textual interpretable model to obtain and then explain the medical diagnosis on Xray images [472]. The employed measures for the image captioning are BLEU, ROUGE, METEOR, and CIDEr, whereas, the sileancy maps are also computed to further elaborate the results. The inter-annotator agreement is computed in terms of Fleiss'

Table 8.2 The summarised results of the studies evaluating the XAI results using different methods

Application	**Objective**	**DL Model**	**XAI Model**	**Performance Scores**	**Year**	**Ref**
Fundus Photography	Glaucoma Detection	CNN	Attention Maps	CC: 0.934 Veriance: 0.0032	2019	[460]
Gastrointestinal	Polyps Classification	ResNet	Grad-CAM	IoU: Grad-CAM: 0.24 Guided Grad-CAM: 0.47 Guided Grad-CAM with Boxes: 0.55	2017	[463]
Skin and Chest	Multiple Disease Classification	ResNet	CAM	Overlapping	2020	[459]
Histology	Tumour Tissue Discrimination	GoogLeNet	Layer-wise Relevance Propagation (LRP)	AUC: 94%	2020	[466]
Cardiovescular	Coronary Artery Atherosclerosis Detection	3D CNN	Grad-CAM	GT and annotation Overlapping. Pixel Level overlap Score. Dice: 0.58 Accuracy: 0.63 Sensitivity: 0.77	2020	[458]
Brain	Multiple Lesion Detection	CNN	CAM based Methods Trainable Attentions	Avg AUC: 72.0 + /- 13.3	2020	[467]
Brain	Multiple Sclerosis Diagnosis	CNN	LRP	Lesion Relevance Score: Accuracy: 96.08%		[457]
Cardiovascular	Stenosis Detection and Classification	Inception V3 LSTM	Grad-CAM	Sensitivity (IoU): 0.72 Mean Square Error (MSE): 69.6	2019	[445]
Brain	Alzheimer's Disease Detection	CNN	LRP Guided Backpropagation	Relevance per Brain Area, e.g., relevance density or relevance gain. >90%	2019	[461]
Brain	Region Discrimination	Pretrained CNN	Prediction Difference Analysis	Max IoU: 54.6% Mean IoU: 43.5%	2020	[465]
Brain	Fetal Brain Aabnormality Detection	CNN	Grad-CAM	Avg Mean IoU: 0.497 Std of IoU: 0.126	2020	[462]
Brain	Alzheimer's Disease Detection	GAN	CAM	Normalised Cross Corelation Mean and Std for CAM: 0.48, 0.04 Mean and Std for VA-GAN: 0.94, 0.07	2018	[464]
Chest	Diagnostic Justification of DL	Pretrained Nets and GAN	Image Captioning	Justification: 2.39 Understanding: 2.45 Agreement: 0.88 Human certainty: 3.75	2019	[469]
Chest	Medical Report Generation	CNN	CIDEr ROUGE BLEU	CIDEr: 0.280 ROUGE-L: 0.339 B1-B2: 0.482, 0.325 B3-B4: 0.226, 0.162 Hit (%): 57.425	2019	[470]

Application	Objective	DL Model	XAI Model	Performance Scores	Year	Ref
Chest	Medical Report Generation	CNN LSTM	CIDEr ROUGE BLEU METEOR	B1-B2: 37.40, 22.41 B3-B4: 15.27, 10.99 CIDEr: 35.97 METEOR: 16.35 ROUGE: 30.76	2019	[471]
Bladder	Medical Image Diagnosis	CNN LSTM	CIDEr ROUGE BLEU METEOR	B1-B2: 91.2, 82.9 B3-B4: 75.0, 76.7 M: 39.6 R: 70.1 C: 2.04	2017	[472]
Chest	Pathologies Location	DenseNet-121 LSTM MLP	BLEU	Atelectasis: 0.61 Effusion: 0.59 Pneumonia: 0.45 Ptx: 0.27	2019	[473]
Chest	Medical Report Generation	CNN LSTM	CIDEr ROUGE BLEU METEOR	B1-B2: 0.517, 0.386 B3-B4: 0.306, 0.247 M: 0.217 R: 0.447 C: 0.327	2018	[474]
Misc	XAI Evaluation	ResNet	LIME SHAP Gradients GSInquire	LIME: 38.05% 35.12% SHAP: 44.15% 40.24% Gradients: 51.22% 47.80% GSInquire: 76.10% 50.73%	2019	[468]

kappa which resulted in 0.33, 0.42 and 0.55 respectively for three different experiments. The additional human evaluation scores on the scale of 1-4 for the parameters Justification, Understanding, Agreement, and Human Certainty reached 2.39, 2.45, 0.88, and 3.75 respectively.

Numerous other articles applied the same performance measuring and evaluation metrics while generating the textual explanation and reporting the diagnostic performances on medical images. These articles studied the DL models for chest Xray and histology images and yielded the performances for CIDEr: 0.280, ROUGE-L: 0.339, BLEU-1: 0.482, BLEU-2: 0.325, BLEU-3: 0.226, BLEU-4: 0.162, and Hit (%): 57.425, respectively by Li et al. [471]; BLEU-1: 37.40, BLEU-2: 22.41, BLEU-3: 15.27, BLEU-4: 10.99, CIDEr: 35.97, METEOR: 16.35, ROUGE: 30.76, respectively by Singh et al. [470]; BLEU-1: 91.2, BLEU-2: 82.9, BLEU-3: 75.0, BLEU-4: 76.7, METEOR: 39.6, ROUGE: 70.1, CIDEr: 2.04, respectively by Zhang et al. [474]; BLEU-1: 0.684, BLEU-2: 0.610, BLEU-3: 0.542, BLEU-4: 0.477, and the IoU for bounding boxes with parameters Atelectasis: 0.61, Effusion: 0.59, Pneumonia: 0.45, Ptx: 0.27, respectively by Roding et al. [473]; and finally BLEU-1: 0.517, BLEU-2:

0.386, BLEU-3: 0.306, BLEU-4: 0.247, METEOR: 0.217, ROUGE: 0.447, CiDEr: 0.327, respectively by Jing et al. [469].

Chapter 9

Conclusion

The objective of the thesis was to design, develop, and validate interpretable and transparent intelligent clinical decision support systems based on DL architectures. The devised systems were sought to be transparent and interpretable on the accounts of mathematical and perceptual explainable techniques. The novel intelligent systems were aimed to assist the medical experts and physicians in the CAD systems and surgical procedures. Such intelligent systems have been designed, developed, and validated with the novel DL techniques and the results are further interpreted with several XAI models. The developed interpretable diagnostic frameworks offer wide range of applications and can be extended to several clinical scenarios. The devised intelligent systems are compared with the state-of-the-art approaches already discussed in the literature. The applicability of the proposed pipelines has also been validated with the assistance of physicians and the domain experts where required.

Conclusively, this thesis has presented the applications of DL for classification, segmentation, and identification tasks and incorporated the XAI methods to increase the interpretability and transparency of CNN models. Through the conceived studies, it has been demonstrated that DL can be a powerful tool for intelligent imaging systems to support the clinicians and physicians in the routine medical tasks. The conjunction of XAI methods help to improve the understanding of DL models and their decision-making processes which builds the necessary trust of medical domain on DL. In light of these contexts, below the concise but independent remarks on the each of the conducted studies are provided.

Breast cancer is the leading deadly ailment in women, and its inevitable progression has become a major concern for the healthcare industry. However, timely diagnosis can significantly improve the medication and prevent the further expansion of the cancerous regions. As part of the thesis work, a robust visually and mathematically explainable DL framework for multiclass shape classification of tomosynthesis breast lesion using eight pretrained CNN models employing an in-house dataset is proposed. The best fine-tuned model achieved mean AUC values of 98.2% and 96.3% with and without considering the data augmentation, respectively.

Moreover, a novel method for segmentation and identification of vertebrae is introduced. The method yields highly accurate results, with an average multi-class Dice coefficient above 90%, with efficiency and ease of use. The framework utilizes unsupervised learning, therefore, it does not require any training data and only needs minimal input from the user. This method has potential clinical value because it can improve the navigation tools used in minimally invasive spine surgery.

Similarly, another proposed pipeline is the fusion prostate biopsy procedure, which involves segmenting TRUS and MRI images using deformable superellipses and nnU-Net, respectively, and registering the two types of images. This procedure is more reliable and accurate than traditional prostate biopsy, which only takes a few samples from specific areas of the prostate without considering the MRI annotations. The segmentation results for both TRUS and MRI images have a Dice coefficient above 88% and 87%, respectively. The image registration step, which is essential for proper image fusion, has a Dice coefficient above 91% for all cases.

The blackbox nature of the decision-making mechanism of the DL architectures hampers the trust among the clinicians. The XAI techniques uncover the blackbox and hidden nature of the DL and provide useful apprehension of the high-accuracy-yielding DL models. This builds confidence in machine learning in the clinical domain and paves the way towards DL-centered image-guided CAD systems.

Therefore, considering the hypersensitive clinical realm, two families of XAI methods, i.e., perceptive interpretability and mathematical interpretability, were incorporated to visually explain the CNN models' classification performance. The former interpretability method includes Grad-CAM and LIME, which are responsible for visually explaining the experimental outcomes in terms of feature-level contribution towards classification, whereas, the latter method comprises t-SNE and UMAP techniques that portray feature clustering capabilities of the DL architectures. The performances of all models were aligned with the visual and mathematical interpretations, hence developing the necessary trust between the healthcare industry and the DL architectures. The results proved the usability of XAI to understand the mechanism of employed AI models, also in the cases of failures.

Furthermore, concerning the case study about image guided surgical applications of DL, four categories including: 1) Surgical Tools, 2) Surgical Processes, 3) Surgical Surveillance, and 4) Surgical Performance/Assessment have been devised. The key findings include: a) Surgical Tools is most studied topic which comprises Surgical Tool Detection and Surgical Tool Segmentation (45% of total studies), b) CNN is most widely applied DL topology (roughly 54% of total studies), c) the gesture recognition studies incorporate JIGSAWS dataset (around 77% of studies in relevant subcategory), whereas MICCAI datasets are top consideration for detection and segmentation tasks (around 60% of studies in relevant subcategory), d) VGG remains the widely accepted pretrained network especially when available dataset was not large enough, f) the most studied applications appear to be cholecystectomy and prostatectomy, g) for gesture and trajectory applications, suturing task is frequently studied application area, h) the fusion of kinematic data with image data yields better performance.

In the healthcare domain, it is imperative to provide clear explanations for the outcomes of DL methods applied to medical images. However, simply visualizing the top contributing features and highlighting important regions on images is not enough to make a DL model completely interpretable. It is therefore necessary to have both qualitative and quantitative measures to evaluate the explanations provided by XAI techniques to build the trust of AI in healthcare industry. The second case study performed during this thesis work examines and investigates such evaluation measures and metrics for XAI, including the quality and types of explanations on medical data.

Unlike other domains, there exists no single generalised quantitative and qualitative method to evaluate the outcomes of XAI in medical imaging domain. Therefore, the future work of the thesis aims to develop and validate a generic pipeline that can be incorporated to explain and evaluate the results of XAI methods. This will not only build the trust among clinicians and DL techniques, but will also abolish the barricade towards completely AI supported autonomous systems.

My Publications

1. Sardar Mehboob Hussain, Antonio Brunetti, Giuseppe Lucarelli, Riccardo Memeo, Vitoantonio Bevilacqua, and Domenico Buongiorno. *Deep Learning based Image Processing for Robot Assisted Surgery: A Systematic Literature Survey.* IEEE Access (2022). doi: 10.1109/ACCESS.2022.3223704.

2. Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua and Antonio Brunetti. *Shape based Breast Lesion Classification using Digital Tomosynthesis Images: the role of Explainable Artificial Intelligence*. Appl. Sci. 2022, 12, 6230. doi: 10.3390/app12126230.

3. Nicola Altini, Antonio Brunetti, Valeria Pia Napoletano, Francesca Girardi, Emanuela Allegretti, Sardar Mehboob Hussain, Gioacchino Brunetti, Vito Triggiani, Vitoantonio Bevilacqua and Domenico Buongiorno. *A Fusion Biopsy Framework for Prostate Cancer Based on Deformable Superellipses and nnU-Net.* Bioengineering 2022, 9, 343. doi: 10.3390/bioengineering9080343.

4. Nicola Altini, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibilano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, Ilaria Sabina Tatò, Gioacchino Brunetti, Vito Triggiani and Vitoantonio Bevilacqua. *Segmentation and Identification of Vertebrae in CT Scans Using CNN, k-Means Clustering and k-NN.* Informatics 2021, 8, 40. doi: 10.3390/informatics8020040.

5. Domenico Buongiorno, Michela Prunella, Stefano Grossi, Sardar Mehboob Hussain, Alessandro Rennola, Nicola Longo, Giovanni Di Stefano, Vitoantonio Bevilacqua, and Antonio Brunetti. *Inline defective laser weld identification by processing thermal image sequences with machine and deep learning techniques.* Applied Sciences 12, no. 13 (2022): 6455. doi.org/10.3390/app12136455

References

[1] Yen-Wei Chen and Lakhmi C Jain. *Deep Learning in Healthcare*. Springer, 2020.

[2] Gobert Lee and Hiroshi Fujita. *Deep learning in medical image analysis: challenges and applications*, volume 1213. Springer, 2020.

[3] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[4] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[5] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47, 2018.

[6] Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11), 2020.

[7] Gunjan Chugh, Shailender Kumar, and Nanhay Singh. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, 13(6): 1451–1470, 2021.

[8] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.

[9] Zahra Rezaei. A review on image-based approaches for breast cancer detection, segmentation, and classification. *Expert Systems with Applications*, 182:115204, 2021.

[10] Essam H Houssein, Marwa M Emam, Abdelmgeid A Ali, and Ponnuthurai Nagaratnam Suganthan. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, 167:114161, 2021.

[11] Gitanjali Wadhwa and Amandeep Kaur. Various image modalities used in computer-aided diagnosis system for detection of breast cancer using machine learning techniques: A systematic review. *Soft Computing and Signal Processing*, pages 281–292, 2022.

[12] Sameera V Mohd Sagheer and Sudhish N George. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61:102036, 2020.

[13] Yingjie Tian and Saiji Fu. A descriptive framework for the field of deep learning applications in medical images. *Knowledge-Based Systems*, 210:106445, 2020.

[14] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.

[15] Wei Wang, Yujing Yang, Xin Wang, Weizheng Wang, and Ji Li. Development of convolutional neural network and its application in image classification: a survey. *Optical Engineering*, 58(4):040901, 2019.

[16] David C Birkhoff, Anne Sophie HM van Dalen, and Marlies P Schijven. A review on the current applications of artificial intelligence in the operating room. *Surgical Innovation*, page 1553350621996961, 2021.

[17] Thomas M Ward, Pietro Mascagni, Yutong Ban, Guy Rosman, Nicolas Padoy, Ozanan Meireles, and Daniel A Hashimoto. Computer vision in surgery. *Surgery*, 169(5): 1253–1256, 2021.

[18] Mecit Can Emre Simsekler, Clarence Rodrigues, Abroon Qazi, Samer Ellahham, and Al Ozonoff. A comparative study of patient and staff safety evaluation using tree-based machine learning algorithms. *Reliability Engineering & System Safety*, 208:107416, 2021.

[19] Jennifer L Fencl, Carrie Willoughby, and Katrina Jackson. Just culture: the foundation of staff safety in the perioperative environment. *AORN journal*, 113(4):329–336, 2021.

[20] Aasia Rehman, Muheet Ahmed Butt, and Majid Zaman. A survey of medical image analysis using deep learning approaches. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1334–1342. IEEE, 2021.

[21] Nishant Kumar and Martin Raubal. Applications of deep learning in congestion detection, prediction and alleviation: A survey. *Transportation Research Part C: Emerging Technologies*, 133:103432, 2021.

[22] Hoangminh Huynhnguyen and Ugo A Buy. Toward gesture recognition in robot-assisted surgical procedures. In *2020 2nd International Conference on Societal Automation (SA)*, pages 1–4. IEEE, 2021.

[23] Zeynettin Akkus, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D Weston, Kenneth A Philbrick, and Bradley J Erickson. A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology*, 16(9):1318–1328, 2019.

[24] Yu Ming, Yang Cheng, Yuan Jing, Li Liangzhe, Yang Pengcheng, Zhang Guang, and Chen Feng. Surgical skills assessment from robot assisted surgery video data. In *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, pages 392–396. IEEE, 2021.

[25] Yang Liu, Jie Jiang, and Jiahao Sun. Hand pose estimation from rgb images based on deep learning: A survey. In *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*, pages 82–89. IEEE, 2021.

[26] Yi Wang, Haoran Dou, Xiaowei Hu, Lei Zhu, Xin Yang, Ming Xu, Jing Qin, Pheng-Ann Heng, Tianfu Wang, and Dong Ni. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. *IEEE transactions on medical imaging*, 38(12):2768–2778, 2019.

[27] Francisco Luongo, Ryan Hakim, Jessica H Nguyen, Animashree Anandkumar, and Andrew J Hung. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery*, 169(5):1240–1244, 2021.

[28] Zonghe Chua, Anthony M Jarc, and Allison M Okamura. Toward force estimation in robot-assisted surgery using deep learning with vision and robot state. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12335–12341. IEEE, 2021.

[29] Intutive Surgical Annual Report 2020. https://www.isrg.intuitive.com/. Last Accessed: [16-05-2022].

[30] Jaydeep H Palep. Robotic assisted minimally invasive surgery. *Journal of minimal access surgery*, 5(1):1, 2009.

[31] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[32] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[33] Mehmet A Gulum, Christopher M Trombley, and Mehmed Kantardzic. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10):4573, 2021.

[34] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

[36] Akm Ashiquzzaman, Sung Min Oh, Dongsu Lee, Jihoon Lee, and Jinsul Kim. Context-aware deep convolutional neural network application for fire and smoke detection in virtual environment for surveillance video analysis. In *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, pages 459–467. Springer, 2021.

[37] Jigang Tang, Songbin Li, and Peng Liu. A review of lane detection methods based on deep learning. *Pattern Recognition*, 111:107623, 2021.

[38] Ramona Magno, Leandro Rocchi, Riccardo Dainelli, Alessandro Matese, Salvatore Filippo Di Gennaro, Chi-Farn Chen, Nguyen-Thanh Son, and Piero Toscano. Agroshadow: A new sentinel-2 cloud shadow detection tool for precision agriculture. *Remote Sensing*, 13(6):1219, 2021.

[39] Henda Boudegga, Yaroub Elloumi, Mohamed Akil, Mohamed Hedi Bedoui, Rostom Kachouri, and Asma Ben Abdallah. Fast and efficient retinal blood vessel segmentation method based on deep learning network. *Computerized Medical Imaging and Graphics*, 90:101902, 2021.

[40] DB Kopans. Mammography, breast imaging. *JB Lippincott Company, Philadelphia*, 30:34–59, 1989.

[41] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss cnn. In *Deep learning and data labeling for medical applications*, pages 121–129. Springer, 2016.

[42] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[43] Edward A Sickles, Carl J D'Orsi, Lawrence W Bassett, Catherine M Appleton, Wendie A Berg, Elizabeth S Burnside, et al. Acr bi-rads® atlas, breast imaging reporting and data system. *Reston, VA: American College of Radiology*, pages 39–48, 2013.

[44] Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua, and Antonio Brunetti. Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence. *Applied Sciences*, 12(12):6230, 2022.

[45] Nicola Altini, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibilano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, et al. Segmentation and identification of vertebrae in ct scans using cnn, k-means clustering and k-nn. In *Informatics*, volume 8, page 40. Multidisciplinary Digital Publishing Institute, 2021.

[46] Nicola Altini, Antonio Brunetti, Valeria Pia Napoletano, Francesca Girardi, Emanuela Allegretti, Sardar Mehboob Hussain, Gioacchino Brunetti, Vito Triggiani, Vitoantonio

Bevilacqua, and Domenico Buongiorno. A fusion biopsy framework for prostate cancer based on deformable superellipses and nnu-net. *Bioengineering*, 9(8):343, 2022.

[47] Sardar Mehboob Hussain, Antonio Brunetti, Giuseppe Lucarelli, Riccardo Memeo, Vitoantonio Bevilacqua, and Domenico Buongiorno. Deep learning based image processing for robot assisted surgery: A systematic literature survey. *IEEE Access*, 2022.

[48] Biswas Mainak, Kuppili Venkatanareshbabu, Saba Luca, Reddy Edla Damodar, Cuadrado-Godia Elisa, Marinhoe Rui Tato, Nicolaides Andrew, et al. State-of-the-art review on deep learning in medical imaging. *Frontiers in Bioscience-Landmark*, 24 (3):380–406, 2019.

[49] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[50] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, 127:104065, 2020.

[51] Shagun Sharma and Kalpna Guleria. Deep learning models for image classification: comparison and applications. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1733–1738. IEEE, 2022.

[52] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42(11):1–13, 2018.

[53] Simone Vicini, Chandra Bortolotto, Marco Rengo, Daniela Ballerini, Davide Bellini, Iacopo Carbone, Lorenzo Preda, Andrea Laghi, Francesca Coppola, and Lorenzo Faggioni. A narrative review on current imaging applications of artificial intelligence and radiomics in oncology: Focus on the three most common cancers. *La radiologia medica*, pages 1–18, 2022.

[54] DR Sarvamangala and Raghavendra V Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1):1–22, 2022.

[55] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[56] Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023.

[57] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, 3:19–48, 2010.

[58] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[59] Xiaomao Li and Qi Tang. Introduction to artificial intelligence and deep learning with a case study in analyzing electronic health records for drug development. In *Real-World Evidence in Drug Development and Evaluation*, pages 151–172. Chapman and Hall/CRC, 2021.

[60] Aristomenis S Lampropoulos and George A Tsihrintzis. Machine learning paradigms. *Appl. Recomm. Syst. Switz. Springer Intern. Publ*, 2015.

[61] Aboul Ella Hassanien et al. *Machine learning paradigms: Theory and application*. Springer, 2019.

[62] Leke Zajmi, Falah YH Ahmed, and Adam Amril Jaharadak. Concepts, methods, and performances of particle swarm optimization, backpropagation, and neural networks. *Applied Computational Intelligence and Soft Computing*, 2018, 2018.

[63] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[64] Congmin Yang, Zijian Zhao, and Sanyuan Hu. Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature. *Computer Assisted Surgery*, 25(1):15–28, 2020.

[65] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[66] Li Li, Miloš Doroslovački, and Murray H Loew. Approximating the gradient of cross-entropy loss function. *IEEE Access*, 8:111626–111635, 2020.

[67] Open source machine learning framework by Google Brains. https://www.tensorflow.org/. Last Accessed: [16-05-2022].

[68] Meta-AI. PyTorch Transforms. https://pytorch.org/vision/stable/transforms.html, 2022. [Online; accessed 15-June-2022].

[69] Numeric computing environment developed by MathWorks. https://www.mathworks.com/products/matlab.html. Last Accessed: [16-05-2022].

[70] GPU based NVIDIA Caffe of Berkeley Vision and Learning Center. https://ngc.nvidia.com/catalog/containers/nvidia:caffe. Last Accessed: [16-05-2022].

[71] Keras library by Google Inc. https://keras.io/. Last Accessed: [16-05-2022].

[72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[73] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[75] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[76] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[77] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[78] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[79] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[80] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[81] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[82] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[83] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[84] Arjun Panesar. *Machine learning and AI for healthcare*. Springer, 2019.

[85] Mohammed Yousef Shaheen. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*, 2021.

[86] Truman Cheng, Weibing Li, Wing Yin Ng, Yisen Huang, Jixiu Li, Calvin Sze Hang Ng, Philip Wai Yan Chiu, and Zheng Li. Deep learning assisted robotic magnetic anchored and guided endoscope for real-time instrument tracking. *IEEE Robotics and Automation Letters*, 6(2):3979–3986, 2021.

[87] Ahmed Ezzat, Alexandros Kogkas, Josephine Holt, Rudrik Thakkar, Ara Darzi, and George Mylonas. An eye-tracking based robotic scrub nurse: proof of concept. *Surgical Endoscopy*, pages 1–11, 2021.

[88] Keitaro Yoshida, Ryo Hachiuma, Hisako Tomita, Jingjing Pan, Kris Kitani, Hiroki Kajita, Tetsu Hayashida, and Maki Sugimoto. Spatiotemporal video highlight by neural network considering gaze and hands of surgeon in egocentric surgical videos. *Journal of Medical Robotics Research*, page 2141001, 2021.

[89] Beatrice van Amsterdam, Matthew Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery: a review. *IEEE Transactions on Biomedical Engineering*, 2021.

[90] Ugo Boggi, Fabio Vistoli, and Gabriella Amorese. Twenty years of robotic surgery: a challenge for human limits, 2021.

[91] A Kanakatte, K Seemakurthy, J Gubbi, J Saha, A Ghose, and B Purushothaman. Surgical smoke dehazing and color reconstruction. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 280–284. IEEE, 2021.

[92] Md Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis*, 70:101994, 2021.

[93] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamin Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014.

[94] MICCAI Annual Challenge Datasets. http://www.miccai.org/. Last Accessed: [16-05-2022].

[95] Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *arXiv preprint arXiv:1808.08180*, 2018.

[96] Duygu Sarikaya, Jason J Corso, and Khurshid A Guru. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE transactions on medical imaging*, 36(7):1542–1549, 2017.

[97] Naila H Dhanani, Oscar A Olavarria, Karla Bernardi, Nicole B Lyons, Julie L Holihan, Michele Loor, Alex B Haynes, and Mike K Liang. The evidence behind robot-assisted abdominopelvic surgery: a systematic review. *Annals of Internal Medicine*, 174(8): 1110–1117, 2021.

[98] Brigid M Gillespie, Joseph Gillespie, Rhonda J Boorman, Karin Granqvist, Johan Stranne, and Annette Erichsen-Andersson. The impact of robotic-assisted surgery on team performance: a systematic mixed studies review. *Human factors*, 63(8):1352–1379, 2021.

[99] Junyu Li, Yanming Fang, Zhao Jin, Yuchen Wang, and Miao Yu. The impact of robot-assisted spine surgeries on clinical outcomes: A systemic review and meta-analysis. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 16(6):1–14, 2020.

[100] Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021.

[101] Erim Yanik, Xavier Intes, Uwe Kruger, Pingkun Yan, David Diller, Brian Van Voorst, Basiel Makled, Jack Norfleet, and Suvranu De. Deep neural networks for the assessment of surgical skills: A systematic review. *The Journal of Defense Modeling and Simulation*, 19(2):159–171, 2022.

[102] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):1–18, 2020.

[103] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018.

[104] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017.

[105] Irene Rivas-Blanco, Carlos J Pérez-Del-Pulgar, Isabel García-Morales, and Víctor F Muñoz. A review on deep learning in minimally invasive surgery. *IEEE Access*, 9: 48658–48678, 2021.

[106] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, page 260, 2021.

[107] Marvin P Fried, Jonathan Kleefield, Harsha Gopal, Edward Reardon, Bryan T Ho, and Frederick A Kuhn. Image-guided endoscopic surgery: results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system. *The Laryngoscope*, 107(5):594–601, 1997.

[108] Austin Reiter, Peter K Allen, and Tao Zhao. Articulated surgical tool detection using virtually-rendered templates. In *Computer assisted radiology and surgery (CARS)*, pages 1–8, 2012.

[109] Robert Elfring, Matías de la Fuente, and Klaus Radermacher. Assessment of optical localizer accuracy for computer aided surgery systems. *Computer Aided Surgery*, 15 (1-3):1–12, 2010.

[110] Yipeng Hu, Hashim Uddin Ahmed, Clare Allen, Doug Pendsé, Mahua Sahu, Mark Emberton, David Hawkes, and Dean Barratt. Mr to ultrasound image registration for guiding prostate biopsy and interventions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 787–794. Springer, 2009.

[111] Apiwat Boonkong, Daranee Hormdee, Suphachoke Sonsilphong, and Kovit Khampitak. Surgical instrument detection for laparoscopic surgery using deep learning. In *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1–4. IEEE, 2022.

[112] Kaustuv Mishra, Rachana Sathish, and Debdoot Sheet. Tracking of retinal microsurgery tools using late fusion of responses from convolutional neural network over pyramidally decomposed frames. In *International Conference on Computer Vision, Graphics, and Image processing*, pages 358–366. Springer, 2016.

[113] Zhaorui Chen, Zijian Zhao, and Xiaolin Cheng. Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context. In *2017 Chinese Automation Congress (CAC)*, pages 2711–2714. IEEE, 2017.

[114] Bareum Choi, Kyungmin Jo, Songe Choi, and Jaesoon Choi. Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1756–1759. Ieee, 2017.

[115] Thomas Probst, Kevis-Kokitsi Maninis, Ajad Chhatkuli, Mouloud Ourak, Emmanuel Vander Poorten, and Luc Van Gool. Automatic tool landmark detection for stereo vision in robot-assisted retinal surgery. *IEEE Robotics and Automation Letters*, 3(1):612–619, 2017.

[116] 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.

[117] Neha Banerjee, Rachana Sathish, and Debdoot Sheet. Deep neural architecture for localization and tracking of surgical tools in cataract surgery. In *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, pages 31–38. Springer, 2019.

[118] Ihsan Ullah, Philip Chikontwe, and Sang Hyun Park. Guidewire tip tracking using u-net with shape and motion constraints. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 215–217. IEEE, 2019.

[119] Mingchuan Zhou, Xijia Wang, Jakob Weiss, Abouzar Eslami, Kai Huang, Mathias Maier, Chris P Lohmann, Nassir Navab, Alois Knoll, and M Ali Nasseri. Needle localization for robot-assisted subretinal injection based on deep learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8727–8732. IEEE, 2019.

[120] Zijian Zhao, Tongbiao Cai, Faliang Chang, and Xiaolin Cheng. Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade. *Healthcare technology letters*, 6(6):275, 2019.

[121] Liang Qiu, Changsheng Li, and Hongliang Ren. Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network. *Healthcare technology letters*, 6(6):159–164, 2019.

[122] Kyungmin Jo, Yuna Choi, Jaesoon Choi, and Jong Woo Chung. Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. *Applied Sciences*, 9(14):2865, 2019.

[123] Yuying Liu, Zijian Zhao, Faliang Chang, and Sanyuan Hu. An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. *IEEE Access*, 8:78193–78201, 2020.

[124] Lena Guinot, Ryosuke Tsumura, Shun Inoue, and Hiroyasu Iwata. Development of a needle deflection detection system for a ct guided robot. In *2020 IEEE/SICE International Symposium on System Integration (SII)*, pages 34–38. IEEE, 2020.

[125] Lingtao Yu, Pengcheng Wang, Yusheng Yan, Yongqiang Xia, and Wei Cao. Massd: Multi-scale attention single shot detector for surgical instruments. *Computers in Biology and Medicine*, 123:103867, 2020.

[126] Ji Woong Kim, Changyan He, Muller Urias, Peter Gehlbach, Gregory D Hager, Iulian Iordachita, and Marin Kobilarov. Autonomously navigating a surgical tool inside the eye by learning from demonstration. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7351–7357. IEEE, 2020.

[127] Tongbiao Cai and Zijian Zhao. Convolutional neural network-based surgical instrument detection. *Technology and Health Care*, 28(S1):81–88, 2020.

[128] Ikjong Park, Hong Kyun Kim, Wan Kyun Chung, and Keehoon Kim. Deep learning based real-time oct image segmentation and correction for robotic needle insertion systems. *IEEE Robotics and Automation Letters*, 5(3):4517–4524, 2020.

[129] Shuai Yin and AS Yuschenko. Object recognition of the robotic system with using a parallel convolutional neural network. In *Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms*, pages 3–11. Springer, 2020.

[130] Atsushi Nakazawa, Kanako Harada, Mamoru Mitsuishi, and Pierre Jannin. Real-time surgical needle detection using region-based convolutional neural networks. *International journal of computer assisted radiology and surgery*, 15(1):41–47, 2020.

[131] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical image analysis*, 59:101572, 2020.

[132] Neil Sachdeva, Misha Klopukh, Rachel St Clair, and William Edward Hahn. Using conditional generative adversarial networks to reduce the effects of latency in robotic telesurgery. *Journal of Robotic Surgery*, pages 1–7, 2020.

[133] Yu Yang, Zijian Zhao, Pan Shi, and Sanyuan Hu. An efficient one-stage detector for real-time surgical tools detection in robot-assisted surgery. In *Annual Conference on Medical Image Understanding and Analysis*, pages 18–29. Springer, 2021.

[134] Sue Min Cho, Young-Gon Kim, Jinhoon Jeong, Inhwan Kim, Ho-jin Lee, and Namkug Kim. Automatic tip detection of surgical instruments in biportal endoscopic spine surgery. *Computers in Biology and Medicine*, 133:104384, 2021.

[135] Shihang Chen, Yanping Lin, Zhaojun Li, Fang Wang, and Qixin Cao. Automatic and accurate needle detection in 2d ultrasound during robot-assisted needle insertion process. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2021.

[136] Yan Wang, Qiyuan Sun, Guodong Sun, Lin Gu, and Zhenzhong Liu. Object detection of surgical instruments based on yolov4. In *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 578–581. IEEE, 2021.

[137] Joonmyeong Choi, Sungman Cho, Jong Woo Chung, and Namkug Kim. Video recognition of simple mastoidectomy using convolutional neural networks: Detection and segmentation of surgical tools and anatomical regions. *Computer Methods and Programs in Biomedicine*, 208:106251, 2021.

[138] Junjun Pan, Dongfang Yu, Ranyang Li, Xin Huang, Xinliang Wang, Wenhao Zheng, Bin Zhu, and Xiaoguang Liu. Multi-modality guidance based surgical navigation for percutaneous endoscopic transforaminal discectomy. *Computer Methods and Programs in Biomedicine*, 212:106460, 2021.

[139] Akito Nakano and Kouki Nagamune. A development of robotic scrub nurse system-detection for surgical instruments using faster region-based convolutional neural network–. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 26(1): 74–82, 2022.

[140] Yisen Huang, Jian Li, Xue Zhang, Ke Xie, Jixiu Li, Yue Liu, Calvin Sze Hang Ng, Philip Wai Yan Chiu, and Zheng Li. A surgeon preference-guided autonomous instrument tracking method with a robotic flexible endoscope based on dvrk platform. *IEEE Robotics and Automation Letters*, 7(2):2250–2257, 2022.

[141] Jani Koskinen, Mastaneh Torkamani-Azar, Ahmed Hussein, Antti Huotarinen, and Roman Bednarik. Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery. *Computers in Biology and Medicine*, 141: 105121, 2022.

[142] Ling Li, Xiaojian Li, Shuai Ding, Zhao Fang, Mengya Xu, Hongliang Ren, and Shanlin Yang. Sirnet: Fine-grained surgical interaction recognition. *IEEE Robotics and Automation Letters*, 7(2):4212–4219, 2022.

[143] Karim Botros, Mohammad Alkhatib, David Folio, and Antoine Ferreira. Fully automatic and real-time microrobot detection and tracking based on ultrasound imaging using deep learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9763–9768. IEEE, 2022.

[144] Yuelin Zou, Bo Guan, Jianchang Zhao, Shuxin Wang, Xinan Sun, and Jianmin Li. Robotic-assisted automatic orientation and insertion for bronchoscopy based on image guidance. *IEEE Transactions on Medical Robotics and Bionics*, 4(3):588–598, 2022.

[145] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. In *International Workshop on Machine Learning in Medical Imaging*, pages 566–573. Springer, 2019.

[146] SM Kamrul Hasan and Cristian A Linte. U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7205–7211. IEEE, 2019.

[147] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3373–3378. IEEE, 2017.

[148] Megha Kalia, Tajwar Abrar Aleef, Nassir Navab, and Septimiu E Salcudean. Co-generation and segmentation for generalized surgical instrument segmentation on unlabelled data. *arXiv preprint arXiv:2103.09276*, 2021.

[149] Thomas Kurmann, Pablo Márquez-Neila, Max Allan, Sebastian Wolf, and Raphael Sznitman. Mask then classify: multi-instance segmentation for surgical instruments. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2021.

[150] Xiaoyan Wang, Luyao Wang, Xingyu Zhong, Cong Bai, Xiaojie Huang, Ruiyi Zhao, and Ming Xia. Pai-net: A modified u-net of reducing semantic gap for surgical instrument segmentation. *IET Image Processing*, 15(12):2959–2969, 2021.

[151] Zhongkai Zhang, Benoît Rosa, and Florent Nageotte. Surgical tool segmentation using generative adversarial networks with unpaired training data. *IEEE Robotics and Automation Letters*, 6(4):6266–6273, 2021.

[152] Jiayi Zhang and Xin Gao. Object extraction via deep learning-based marker-free tracking framework of surgical instruments for laparoscope-holder robots. *International Journal of Computer Assisted Radiology and Surgery*, 15(8):1335–1345, 2020.

[153] Dominik Rivoir, Sebastian Bodenstedt, Isabel Funke, Felix von Bechtolsheim, Marius Distler, Jürgen Weitz, and Stefanie Speidel. Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 752–762. Springer, 2020.

[154] Saul Alexis Heredia Perez, Murilo Marques Marinho, Kanako Harada, and Mamoru Mitsuishi. The effects of different levels of realism on the training of cnns with only synthetic images for the semantic segmentation of robotic instruments in a head phantom. *International Journal of Computer Assisted Radiology and Surgery*, 15:1257–1265, 2020.

[155] Fangbo Qin, Shan Lin, Yangming Li, Randall A Bly, Kris S Moe, and Blake Hannaford. Towards better surgical instrument segmentation in endoscopic vision: multi-angle feature aggregation and contour supervision. *IEEE Robotics and Automation Letters*, 5 (4):6639–6646, 2020.

[156] Cheng-Shao Chiang and Chi-Sheng Daniel Shih. Using synthesized data to train deep neural net with few data. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pages 19–25, 2020.

[157] Kateryna Zinchenko and Kai-Tai Song. Autonomous endoscope robot positioning using instrument segmentation with virtual reality visualization. *IEEE Access*, 9:72614–72623, 2021.

[158] Zhen-Liang Ni, Gui-Bin Bian, Guan-An Wang, Xiao-Hu Zhou, Zeng-Guang Hou, Hua-Bin Chen, and Xiao-Liang Xie. Pyramid attention aggregation network for semantic segmentation of surgical instruments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 07, pages 11782–11790, 2020.

[159] Shan Lin, Fangbo Qin, Yangming Li, Randall A Bly, Kris S Moe, and Blake Hannaford. Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2914–2920. IEEE, 2020.

[160] Rocco Moccia, Cristina Iacono, Bruno Siciliano, and Fanny Ficuciello. Vision-based dynamic virtual fixtures for tools collision avoidance in robotic surgery. *IEEE Robotics and Automation Letters*, 5(2):1650–1655, 2020.

[161] Daniil Pakhomov and Nassir Navab. Searching for efficient architecture for instrument segmentation in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–656. Springer, 2020.

[162] Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 784–794. Springer, 2020.

[163] Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters*, 7(2):3858–3865, 2022.

[164] Haibin Wu, Jianbo Zhao, Kaiyang Xu, Yan Zhang, Ruotong Xu, Aili Wang, and Yuji Iwahori. Semantic slam based on deep learning in endocavity environment. *Symmetry*, 14(3):614, 2022.

[165] Zhen-Liang Ni, Xiao-Hu Zhou, Guan-An Wang, Wen-Qian Yue, Zhen Li, Gui-Bin Bian, and Zeng-Guang Hou. Surginet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. *Medical Image Analysis*, 76:102310, 2022.

[166] Tahir Mahmood, Se Woon Cho, and Kang Ryoung Park. Dsrd-net: Dual-stream residual dense network for semantic segmentation of instruments in robot-assisted surgery. *Expert Systems with Applications*, 202:117420, 2022.

[167] Baoru Huang, Anh Nguyen, Siyao Wang, Ziyang Wang, Erik Mayer, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. *IEEE Transactions on Medical Robotics and Bionics*, 4(2):335–338, 2022.

[168] Lei Yang, Yuge Gu, Guibin Bian, and Yanhong Liu. Drr-net: A dense-connected residual recurrent convolutional network for surgical instrument segmentation from endoscopic images. *IEEE Transactions on Medical Robotics and Bionics*, 4(3):696–707, 2022.

[169] Kevin Huang, Digesh Chitrakar, Wenfan Jiang, Isabella Yung, and Yun-Hsuan Su. Surgical tool segmentation with pose-informed morphological polar transform of endoscopic images. *Journal of Medical Robotics Research*, 2022.

[170] Xinan Sun, Yuelin Zou, Shuxin Wang, He Su, and Bo Guan. A parallel network utilizing local features and global representations for segmentation of surgical instruments. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11, 2022.

[171] Suphachoke Sonsilphong, Amornthep Sonsilphong, Daranee Hormdee, and Kovit Khampitak. A development of object detection system based on deep learning approach to support the laparoscope manipulating robot (lmr). In *2022 International Electrical Engineering Congress (iEECON)*, pages 1–4. IEEE, 2022.

[172] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters*, 4(2):2188–2195, 2019.

[173] Jiacheng Wang, Yueming Jin, Liansheng Wang, Shuntian Cai, Pheng-Ann Heng, and Jing Qin. Efficient global-local memory for real-time instrument segmentation of robotic surgical video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 341–351. Springer, 2021.

[174] Shan Lin, Fangbo Qin, Haonan Peng, Randall A Bly, Kris S Moe, and Blake Hannaford. Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video. *IEEE Robotics and Automation Letters*, 6(4):6773–6780, 2021.

[175] Zhen-Liang Ni, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Hu Zhou, Xiao-Liang Xie, and Zhen Li. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9939–9945. IEEE, 2020.

[176] Emanuele Colleoni, Philip Edwards, and Danail Stoyanov. Synthetic and real inputs for tool segmentation in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 700–710. Springer, 2020.

[177] Yanwen Sun, Bo Pan, and Yili Fu. Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery. *IEEE Robotics and Automation Letters*, 6(2):3870–3877, 2021.

[178] Yu-Dong Wu, Xiao-Liang Xie, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Ran Cheng, Sheng Chen, Shi-Qi Liu, and Qiao-Li Wang. Automatic guidewire tip segmentation in 2d x-ray fluoroscopy using convolution neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[179] Jia Yi Lee, Mobarakol Islam, Jing Ru Woh, TS Mohamed Washeem, Lee Ying Clara Ngoh, Weng Kin Wong, and Hongliang Ren. Ultrasound needle segmentation and trajectory prediction using excitation network. *International journal of computer assisted radiology and surgery*, 15(3):437–443, 2020.

[180] Wanghongbo Li and Dashun Que. Research on master-slave isomorphism design and guide wire segmentation of robot for vascular intervention. In *2021 International Conference on Robotics and Control Engineering*, pages 7–12, 2021.

[181] Daniil Pakhomov, Wei Shen, and Nassir Navab. Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8499–8504. IEEE, 2020.

[182] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–667. Springer, 2020.

[183] Xiaojie Gao, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8440–8446. IEEE, 2020.

[184] Jinglu Zhang, Yinyu Nie, Yao Lyu, Xiaosong Yang, Jian Chang, and Jian Jun Zhang. Sd-net: joint surgical gesture recognition and skill assessment. *International Journal of Computer Assisted Radiology and Surgery*, 16(10):1675–1682, 2021.

[185] Danit Itzkovich, Yarden Sharon, Anthony Jarc, Yael Refaely, and Ilana Nisky. Using augmentation to improve the robustness to rotation of deep learning segmentation in robotic-assisted surgical data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5068–5075. IEEE, 2019.

[186] Isabel Funke, Sebastian Bodenstedt, Florian Oehme, Felix von Bechtolsheim, Jürgen Weitz, and Stefanie Speidel. Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–475. Springer, 2019.

[187] Jeffrey Hsu and Shahram Payandeh. Toward tool gesture and motion recognition on a novel minimally invasive surgery robotic system. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 631–636. IEEE, 2006.

[188] Wenjie Wang, Mengling He, Xiaohua Wang, Jianwei Ma, and Huajian Song. Medical gesture recognition method based on improved lightweight network. *Applied Sciences*, 12(13):6414, 2022.

[189] Snigdha Agarwal, Chakka Sai Pradeep, and Neelam Sinha. Temporal surgical gesture segmentation and classification in multi-gesture robotic surgery using fine-tuned features and calibrated ms-tcn. In *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2022.

[190] Beatrice van Amsterdam, Matthew J Clarkson, and Danail Stoyanov. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1380–1386. IEEE, 2020.

[191] Jinglu Zhang, Yinyu Nie, Yao Lyu, Hailin Li, Jian Chang, Xiaosong Yang, and Jian Jun Zhang. Symmetric dilated convolution for surgical gesture recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 409–418. Springer, 2020.

[192] Danit Itzkovich, Yarden Sharon, Anthony Jarc, Yael Refaely, and Ilana Nisky. Generalization of deep learning gesture classification in robotic-assisted surgical data: From dry lab to clinical-like data. *IEEE Journal of Biomedical and Health Informatics*, 26(3): 1329–1340, 2021.

[193] Beatrice Van Amsterdam, Isabel Funke, Eddie Edwards, Stefanie Speidel, Justin Collins, Ashwin Sridhar, John Kelly, Matthew J Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery with multimodal attention. *IEEE Transactions on Medical Imaging*, 2022.

[194] Hongfa Zhao, Jiexin Xie, Zhenzhou Shao, Ying Qu, Yong Guan, and Jindong Tan. A fast unsupervised approach for multi-modality surgical trajectory segmentation. *IEEE Access*, 6:56411–56422, 2018.

[195] Bo Lu, XB Yu, JW Lai, KC Huang, Keith CC Chan, and Henry K Chu. A learning approach for suture thread detection with feature enhancement and segmentation for 3-d shape reconstruction. *IEEE Transactions on Automation Science and Engineering*, 17 (2):858–870, 2019.

[196] Dongyang Cai, Zaiyue Wang, Yajun Liu, Qi Zhang, Xiaoguang Han, and Wenyong Liu. Automatic path planning for navigated pedicle screw surgery based on deep neural network. In *2019 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, pages 62–67. IEEE, 2019.

[197] Qian Li, Zhijiang Du, and Hongjian Yu. Grinding trajectory generator in robot-assisted laminectomy surgery. *International Journal of Computer Assisted Radiology and Surgery*, 16(3):485–494, 2021.

[198] Brijen Thananjeyan, Animesh Garg, Sanjay Krishnan, Carolyn Chen, Lauren Miller, and Ken Goldberg. Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2371–2378. IEEE, 2017.

[199] Adithyavairavan Murali, Animesh Garg, Sanjay Krishnan, Florian T Pokorny, Pieter Abbeel, Trevor Darrell, and Ken Goldberg. Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4150–4157. IEEE, 2016.

[200] Zhenzhou Shao, Hongfa Zhao, Jiexin Xie, Ying Qu, Yong Guan, and Jindong Tan. Unsupervised trajectory segmentation and promoting of multi-modal surgical demonstrations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 777–782. IEEE, 2018.

[201] Jacky Liang, Jeffrey Mahler, Michael Laskey, Pusong Li, and Ken Goldberg. Using dvrk teleoperation to facilitate deep learning of automation tasks for an industrial robot. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, pages 1–8. IEEE, 2017.

[202] Yan Zhao, Yuxin Wang, Jianhua Zhang, Xinke Liu, Youxiang Li, Shuxiang Guo, Xu Yang, and Shunming Hong. Surgical gan: Towards real-time path planning for passive flexible tools in endovascular surgeries. *Neurocomputing*, 2022.

[203] Chen Yao, Jishuai He, Hui Che, Yibin Huang, and Jian Wu. Feature pyramid self-attention network for respiratory motion prediction in ultrasound image guided surgery. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2022.

[204] Daniele De Gregorio, Gianluca Palli, and Luigi Di Stefano. Let's take a walk on superpixels graphs: Deformable linear objects segmentation and model estimation. In *Asian Conference on Computer Vision*, pages 662–677. Springer, 2018.

[205] Qian Li, Zhijiang Du, and Hongjian Yu. Trajectory planning for robot-assisted laminectomy decompression based on ct images. In *IOP Conference Series: Materials Science and Engineering*, volume 768, No. 4, page 042037. IOP Publishing, 2020.

[206] Xiaozhi Qi, Jin Meng, Meng Li, Yuanyuan Yang, Ying Hu, Bing Li, Jianwei Zhang, and Wei Tian. An automatic path planning method of pedicle screw placement based on preoperative ct images. *IEEE Transactions on Medical Robotics and Bionics*, 4(2): 403–413, 2022.

[207] Aleks Attanasio, Chiara Alberti, Bruno Scaglioni, Nils Marahrens, Alejandro F Frangi, Matteo Leonetti, Chandra Shekhar Biyani, Elena De Momi, and Pietro Valdastri. A comparative study of spatio-temporal u-nets for tissue segmentation in surgical robotics. *IEEE Transactions on Medical Robotics and Bionics*, 3(1):53–63, 2021.

[208] Yaqub Jonmohamadi, Yu Takeda, Fengbei Liu, Fumio Sasazawa, Gabriel Maicas, Ross Crawford, Jonathan Roberts, Ajay K Pandey, and Gustavo Carneiro. Automatic segmentation of multiple structures in knee arthroscopy using deep learning. *IEEE Access*, 8:51853–51861, 2020.

[209] Changyeob Shin, Peter Walker Ferguson, Sahba Aghajani Pedram, Ji Ma, Erik P Dutson, and Jacob Rosen. Autonomous tissue manipulation via surgical robot using learning based model predictive control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3875–3881. IEEE, 2019.

[210] Aleks Attanasio, Bruno Scaglioni, Matteo Leonetti, Alejandro F Frangi, William Cross, Chandra Shekhar Biyani, and Pietro Valdastri. Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study. *IEEE Robotics and Automation Letters*, 5(4):6528–6535, 2020.

[211] Maria Antico, Fumio Sasazawa, Yu Takeda, Anjali Tumkur Jaiprakash, Marie-Luise Wille, Ajay K Pandey, Ross Crawford, Gustavo Carneiro, and Davide Fontanarosa. Bayesian cnn for segmentation uncertainty inference on 4d ultrasound images of the femoral cartilage for guidance in robotic knee arthroscopy. *IEEE Access*, 8:223961–223975, 2020.

[212] Ryan J Murphy, Matthew S Moses, Michael DM Kutzer, Gregory S Chirikjian, and Mehran Armand. Constrained workspace generation for snake-like manipulators with applications to minimally invasive surgery. In *2013 IEEE International Conference on Robotics and Automation*, pages 5341–5347. IEEE, 2013.

[213] Yachun Li, Patra Charalampaki, Yong Liu, Guang-Zhong Yang, and Stamatia Giannarou. Context aware decision support in neurosurgical oncology based on an efficient classification of endomicroscopic data. *International journal of computer assisted radiology and surgery*, 13(8):1187–1199, 2018.

[214] Paul Maria Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat P Müller-Stich, Martin Wagner, and Franziska Mathis-Ullrich. Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. *Current Directions in Biomedical Engineering*, 6(1), 2020.

[215] Maria Antico, Fumio Sasazawa, Matteo Dunnhofer, SM Camps, AT Jaiprakash, AK Pandey, Ross Crawford, Gustavo Carneiro, and Davide Fontanarosa. Deep learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy. *Ultrasound in medicine & biology*, 46(2):422–435, 2020.

[216] Yuta Kumazu, Nao Kobayashi, Naoki Kitamura, Elleuch Rayan, Paul Neculoiu, Toshihiro Misumi, Yudai Hojo, Tatsuro Nakamura, Tsutomu Kumamoto, Yasunori Kurahashi, et al. Automated segmentation by deep learning of loose connective tissue fibers to define safe dissection planes in robot-assisted gastrectomy. *Scientific Reports*, 11(1):1–10, 2021.

[217] Aravind Venugopal, Sara Moccia, Simone Foti, Arpita Routray, Robert A MacLachlan, Alessandro Perin, Leonardo S Mattos, Alexander K Yu, Jody Leonardo, Elena De Momi, et al. Real-time vessel segmentation and reconstruction for virtual fixtures for an active handheld microneurosurgical instrument. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2022.

[218] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Berkin Özdemir, Manuel Wiesenfarth, Leonardo Ayala, Jan Odenthal, Samuel Knödler, Karl-Friedrich Kowalewski, Caelan Max Haney, et al. Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hyperspectral imaging. 2022.

[219] Eli Gibson, Maria R Robu, Stephen Thompson, P Eddie Edwards, Crispin Schneider, Kurinchi Gurusamy, Brian Davidson, David J Hawkes, Dean C Barratt, and Matthew J Clarkson. Deep residual networks for automatic segmentation of laparoscopic videos of the liver. In *Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10135, page 101351M. International Society for Optics and Photonics, 2017.

[220] Georges Hattab, Marvin Arnold, Leon Strenger, Max Allan, Darja Arsentjeva, Oliver Gold, Tobias Simpfendörfer, Lena Maier-Hein, and Stefanie Speidel. Kidney edge detection in laparoscopic image data for computer-assisted surgery. *International journal of computer assisted radiology and surgery*, 15(3):379–387, 2020.

[221] Alessandro Casella, Sara Moccia, Chiara Carlini, Emanuele Frontoni, Elena De Momi, and Leonardo S Mattos. Nephcnn: A deep-learning framework for vessel segmentation in nephrectomy laparoscopic videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6144–6149. IEEE, 2021.

[222] M Sadeghi-Goughari, A Mojra, and S Sadeghi. Parameter estimation of brain tumors using intraoperative thermal imaging based on artificial tactile sensing in conjunction with artificial neural network. *Journal of Physics D: Applied Physics*, 49(7):075404, 2016.

[223] Leonardo Tanzi, Pietro Piazzolla, Francesco Porpiglia, and Enrico Vezzetti. Real-time deep learning semantic segmentation during intra-operative surgery for 3d augmented reality assistance. *International Journal of Computer Assisted Radiology and Surgery*, 16(9):1435–1445, 2021.

[224] Yanru Miao, Yu Sun, Shibo Li, Peng Zhang, Yuanyuan Yang, and Ying Hu. Spinal neoplasm image inpainting with deep convolutional neutral networks. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2619–2624. IEEE, 2019.

[225] Stephen G Laws, Spyridon Souipas, Brian L Davies, and Ferdinando Rodriguez y Baena. Toward automated tissue classification for markerless orthopaedic robotic assistance. *IEEE Transactions on Medical Robotics and Bionics*, 2(4):537–540, 2020.

[226] Mark Marsden, Brent W Weyers, Julien Bec, Tianchen Sun, Regina F Gandour-Edwards, Andrew C Birkeland, Marianne Abouyared, Arnaud F Bewley, D Gregory Farwell, and Laura Marcu. Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning. *IEEE Transactions on Biomedical Engineering*, 68(3):857–868, 2020.

[227] Laura J Brattain, Theodore T Pierce, Lars A Gjesteby, Matthew R Johnson, Nancy D DeLosa, Joshua S Werblin, Jay F Gupta, Arinc Ozturk, Xiaohong Wang, Qian Li, et al. Ai-enabled, ultrasound-guided handheld robotic device for femoral vascular access. *Biosensors*, 11(12):522, 2021.

[228] Zheng Li, Xiaofeng Zhang, Lele Ding, Kebin Du, Jun Yan, Matthew TV Chan, William KK Wu, and Shugang Li. Deep learning approach for guiding three-dimensional computed tomography reconstruction of lower limbs for robotically-assisted total knee arthroplasty. *The International Journal of Medical Robotics and Computer Assisted Surgery*, page e2300, 2021.

[229] Elias Eulig, Joscha Maier, Michael Knaup, N Robert Bennett, Klaus Hörndler, Adam S Wang, and Marc Kachelrieß. Deep learning-based reconstruction of interventional tools and devices from four x-ray projections for tomographic interventional guidance. *Medical physics*, 48(10):5837–5850, 2021.

[230] Clement Baumgarten, Y Zhao, Paul Sauleau, Cecile Malrain, Pierre Jannin, and Claire Haegelen. Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation. In *Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 9786, page 97860U. International Society for Optics and Photonics, 2016.

[231] YM Zhao, Edward H Currie, Louis Kavoussi, and Sina Y Rabbany. Laser scanner for 3d reconstruction of a wound's edge and topology. *International Journal of Computer Assisted Radiology and Surgery*, 16(10):1761–1773, 2021.

[232] Qian Li, Zhijiang Du, and Hongjian Yu. Precise laminae segmentation based on neural network for robot-assisted decompressive laminectomy. *Computer Methods and Programs in Biomedicine*, 209:106333, 2021.

[233] Jordina Torrents-Barrena, Rocío López-Velazco, Narcís Masoller, Brenda Valenzuela-Alcaraz, Eduard Gratacós, Elisenda Eixarch, Mario Ceresa, and Miguel Ángel González Ballester. Preoperative planning and simulation framework for twin-to-twin transfusion syndrome fetal surgery. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 184–193. Springer, 2018.

[234] Xiao-Yun Zhou, Jianyu Lin, Celia Riga, Guang-Zhong Yang, and Su-Lin Lee. Real-time 3-d shape instantiation from single fluoroscopy projection for fenestrated stent graft deployment. *IEEE Robotics and Automation Letters*, 3(2):1314–1321, 2018.

[235] Anam Nazir, Muhammad Nadeem Cheema, Bin Sheng, Ping Li, Huating Li, Po Yang, Younhyun Jung, Jing Qin, and David Dagan Feng. Spst-cnn: Spatial pyramid based searching and tagging of liver's intraoperative live views via cnn for minimal invasive surgery. *Journal of biomedical informatics*, 106:103430, 2020.

[236] Giacomo De Rossi, Serena Roin, Francesco Setti, and Riccardo Muradore. A multi-modal learning system for on-line surgical action segmentation. In *2020 International Symposium on Medical Robotics (ISMR)*, pages 132–138. IEEE, 2020.

[237] Yan Zhao, Shuxiang Guo, Yuxin Wang, Jinxin Cui, Youchun Ma, Yuwen Zeng, Xinke Liu, Yuhua Jiang, Youxinag Li, Liwei Shi, et al. A cnn-based prototype method of

unstructured surgical state perception and navigation for an endovascular surgery robot. *Medical & Biological Engineering & Computing*, 57(9):1875–1887, 2019.

[238] Yidan Qin, Max Allan, Joel Burdick, and Mahdi Azizian. Autonomous hierarchical surgical state estimation during robot-assisted surgery through deep neural networks. *IEEE Robotics and Automation Letters*, 2021.

[239] Yidan Qin, Max Allan, Yisong Yue, Joel W Burdick, and Mahdi Azizian. Learning invariant representation of tasks for robust surgical state estimation. *IEEE Robotics and Automation Letters*, 6(2):3208–3215, 2021.

[240] Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2021.

[241] Takuto Mikada, Takahiro Kanno, Toshihiro Kawase, Tetsuro Miyazaki, and Kenji Kawashima. Suturing support by human cooperative robot control using deep learning. *IEEE Access*, 8:167739–167746, 2020.

[242] Masashi Takeuchi, Hirofumi Kawakubo, Kosuke Saito, Yusuke Maeda, Satoru Matsuda, Kazumasa Fukuda, Rieko Nakamura, and Yuko Kitagawa. Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence. *Annals of Surgical Oncology*, pages 1–9, 2022.

[243] Xinpeng Ding and Xiaomeng Li. Exploring segment-level semantics for online phase recognition from surgical videos. *IEEE Transactions on Medical Imaging*, 2022.

[244] Yangxi Li, Yingwei Fan, Chengquan Hu, Fan Mao, Xinran Zhang, and Hongen Liao. Intelligent optical diagnosis and treatment system for automated image-guided laser ablation of tumors. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2147–2157, 2021.

[245] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 385–395. Springer, 2020.

[246] Siddharth Kannan, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Future-state predicting lstm for early surgery type recognition. *IEEE transactions on medical imaging*, 39(3):556–566, 2019.

[247] Toktam Khatibi and Parastoo Dezyani. Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos. *Multimedia Tools and Applications*, 79(41):30111–30133, 2020.

[248] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 364–374. Springer, 2020.

[249] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2016.

[250] Aneeq Zia, Andrew Hung, Irfan Essa, and Anthony Jarc. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–280. Springer, 2018.

[251] ImageNet Large Scale Visual Recognition Challenge. Olga russakovsky, jia deng, hao su, jonathan krause, sanjeev satheesh, sean ma, zhiheng huang, andrej karpathy, aditya khosla, michael bernstein, alexander c. berg, li fei-fei. *Computing Research Repository, Vol. abs/1409.0575*, 2014.

[252] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–221. Springer, 2018.

[253] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.

[254] Dongheon Lee, Hyeong Won Yu, Hyungju Kwon, Hyoun-Joong Kong, Kyu Eun Lee, and Hee Chan Kim. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *Journal of clinical medicine*, 9(6):1964, 2020.

[255] Dandan Zhang, Zicong Wu, Junhong Chen, Anzhu Gao, Xu Chen, Peichao Li, Zhaoyang Wang, Guitao Yang, Benny Lo, and Guang-Zhong Yang. Automatic microsurgical skill assessment based on cross-domain transfer learning. *IEEE Robotics and Automation Letters*, 5(3):4148–4155, 2020.

[256] Roger Smith, Danielle Julian, and Ariel Dubin. Deep neural networks are effective tools for assessing performance during surgical training. *Journal of Robotic Surgery*, pages 1–4, 2021.

[257] Gábor Lajkó, Renáta Nagyné Elek, and Tamás Haidegger. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors*, 21(16):5412, 2021.

[258] Abed Soleymani, Xingyu Li, and Mahdi Tavakoli. A domain-adapted machine learning approach for visual evaluation and interpretation of robot-assisted surgery skills. *IEEE Robotics and Automation Letters*, 7(3):8202–8208, 2022.

[259] Elizebeth Kurian, Jubilant J Kizhakethottam, and Justin Mathew. Deep learning based surgical workflow recognition from laparoscopic videos. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 928–931. IEEE, 2020.

[260] Xueying Shi, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Lrtd: long-range temporal dependency based active learning for surgical workflow recognition. *International Journal of Computer Assisted Radiology and Surgery*, 15(9):1573–1584, 2020.

[261] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 37(5):1114–1126, 2017.

[262] Isabel Funke, Alexander Jenke, Sören Torge Mees, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*, pages 85–93. Springer, 2018.

[263] Hirenkumar Nakawala, Roberto Bianchi, Laura Erica Pescatori, Ottavio De Cobelli, Giancarlo Ferrigno, and Elena De Momi. "deep-onto" network for surgical workflow and context recognition. *International journal of computer assisted radiology and surgery*, 14(4):685–696, 2019.

[264] Hirenkumar Nakawala, Elena De Momi, Roberto Bianchi, Michele Catellani, Ottavio De Cobelli, Pierre Jannin, Giancarlo Ferrigno, and Paolo Fiorini. Toward a neural-symbolic framework for automated workflow analysis in surgery. In *Mediterranean Conference on Medical and Biological Engineering and Computing*, pages 1551–1558. Springer, 2019.

[265] Ella Lan. A novel deep learning architecture by integrating visual simultaneous localization and mapping (vslam) into cnn for real-time surgical video analysis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[266] Claudio S Ravasio, Theodoros Pissas, Edward Bloch, Blanca Flores, Sepehr Jalali, Danail Stoyanov, Jorge M Cardoso, Lyndon Da Cruz, and Christos Bergeles. Learned optical flow for intra-operative tracking of the retinal fundus. *International journal of computer assisted radiology and surgery*, 15(5):827–836, 2020.

[267] Ralf Stauder, Daniel Ostler, Thomas Vogel, Dirk Wilhelm, Sebastian Koller, Michael Kranzfelder, and Nassir Navab. Surgical data processing for smart intraoperative assistance systems. *Innovative surgical sciences*, 2(3):145–152, 2017.

[268] Erica Padovan, Giorgia Marullo, Leonardo Tanzi, Pietro Piazzolla, Sandro Moos, Francesco Porpiglia, and Enrico Vezzetti. A deep learning framework for real-time 3d model registration in robot-assisted laparoscopic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18(3):e2387, 2022.

[269] Igor Artemchuk, Eduard Petlenkov, and Fujio Miyawaki. Neural network based system for real-time organ recognition during surgical operation. *IFAC Proceedings Volumes*, 44(1):6478–6483, 2011.

[270] Liang Li, Pengfei Feng, Hui Ding, and Guangzhi Wang. A preliminary exploration to make stereotactic surgery robots aware of the semantic 2d/3d working scene. *IEEE Transactions on Medical Robotics and Bionics*, 4(1):17–27, 2021.

[271] Silvia Seidlitz, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J Adler, Hannes G Kenngott, Minu Tizabi, et al. Robust deep learning-based semantic organ segmentation in hyperspectral images. *Medical Image Analysis*, page 102488, 2022.

[272] Peichao Li, Xiao-Yun Zhou, Zhao-Yang Wang, and Guang-Zhong Yang. Z-net: An anisotropic 3d dcnn for medical ct volume segmentation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2906–2913. IEEE, 2020.

[273] Peng Li, Xuebin Hou, Le Wei, Guoli Song, and Xingguang Duan. Efficient and low-cost deep-learning based gaze estimator for surgical robot control. In *2018 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 58–63. IEEE, 2018.

[274] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian Jun Zhang. Desmokegcn: generative cooperative networks for joint surgical smoke detection and removal. *IEEE transactions on medical imaging*, 39(5):1615–1625, 2019.

[275] Andreas Leibetseder, Manfred Jürgen Primus, Stefan Petscharnig, and Klaus Schoeffmann. Image-based smoke detection in laparoscopic videos. In *Computer assisted and robotic endoscopy and clinical image-based procedures*, pages 70–87. Springer, 2017.

[276] Yirou Pan, Sophia Bano, Francisco Vasconcelos, Hyun Park, Taikyeong Ted Jeong, and Danail Stoyanov. Desmoke-lap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 17(5):885–893, 2022.

[277] Guillermo Sánchez-Brizuela, Francisco-Javier Santos-Criado, Daniel Sanz-Gobernado, Eusebio de la Fuente-López, Juan-Carlos Fraile, Javier Pérez-Turiel, and Ana Cisnal. Gauze detection and segmentation in minimally invasive surgery video using convolutional neural networks. *Sensors*, 22(14):5180, 2022.

[278] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[279] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 691–699. IEEE, 2018.

[280] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab. The tum lapchole dataset for the m2cai 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*, 2016.

[281] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018.

[282] Shan Lin, Fangbo Qin, Randall A Bly, Kris S Moe, and Blake Hannaford. Uw sinus surgery cadaver/live dataset (uw-sinus-surgery-c/l). *digital.lib.washington.edu*, 2020.

[283] Kyle H Sheetz, Jake Claflin, and Justin B Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA network open*, 3(1):e1918911–e1918911, 2020.

[284] Christopher JD Wallis and Allan S Detsky. Pitfalls of prioritizing cost-effectiveness in the assessment of medical innovation. *Canadian Urological Association Journal*, 12(2): 7, 2018.

[285] Henriette Roscam Abbing. Innovative technologies in healthcare, beware of the pitfalls. *European Journal of Health Law*, 27(1):1–8, 2020.

[286] Douglas C Hague. Benefits, pitfalls, and potential bias in health care ai. *North Carolina medical journal*, 80(4):219–223, 2019.

[287] Dominique Thomas, Brent Medoff, Jennifer Anger, and Bilal Chughtai. Direct-to-consumer advertising for robotic surgery. *Journal of robotic surgery*, 14(1):17–20, 2020.

[288] Rossella Onofrio and Paolo Trucco. A methodology for dynamic human reliability analysis in robotic surgery. *Applied Ergonomics*, 88:103150, 2020.

[289] Qaysar Salih Mahdi, Idris Hadi Saleh, Ghani Hashim, and Ganesh Babu Loganathan. Evaluation of robot professor technology in teaching and business. *Information Technology in Industry*, 9(1):1182–1194, 2021.

[290] Russell Belk. Ethical issues in service robotics and artificial intelligence. *The Service Industries Journal*, 41(13-14):860–876, 2021.

[291] Daniel Schönberger. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*, 27 (2):171–203, 2019.

[292] Justin W Collins, Hani J Marcus, Ahmed Ghazi, Ashwin Sridhar, Daniel Hashimoto, Gregory Hager, Alberto Arezzo, Pierre Jannin, Lena Maier-Hein, Keno Marz, et al. Ethical implications of ai in robotic surgical training: A delphi consensus statement. *European Urology Focus*, 2021.

[293] Sara Gerke, Timo Minssen, and Glenn Cohen. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare*, pages 295–336. Elsevier, 2020.

[294] Tamara Bonaci, Jeffrey Herron, Tariq Yusuf, Junjie Yan, Tadayoshi Kohno, and Howard Jay Chizeck. To make a robot secure: An experimental analysis of cyber security threats against teleoperated surgical robots. *arXiv preprint arXiv:1504.04339*, 2015.

[295] Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (ai) and autonomous robotic surgery. *The international journal of medical robotics and computer assisted surgery*, 15(1):e1968, 2019.

[296] Roger Collier. Nhs ransomware attack spreads worldwide, 2017.

[297] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[298] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020.

[299] Quande Liu, Qi Dou, and Pheng Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

[300] Andrey Fedorov, Paul L Nguyen, Kemal Tuncali, and Clare Tempany. Annotated mri and ultrasound volume images of the prostate, 2015. URL https://doi.org/10.5281/zenodo.16396.

[301] World Health Organization. Worldwide cancer data. *World Cancer Research Fund*, pages 7–12, 2018. URL https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data%0Ainternal-pdf://0.0.1.154/worldwide-cancer-data.html.

[302] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2): 63, 2019.

[303] Soumya Ghose, Arnau Oliver, Robert Martí, Xavier Lladó, Joan C Vilanova, Jordi Freixenet, Jhimli Mitra, Désiré Sidibé, and Fabrice Meriaudeau. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. *Computer methods and programs in biomedicine*, 108(1):262–287, 2012.

[304] Konstantinos Devetzis, Francesca Kum, and Richard Popert. Recent advances in systematic and targeted prostate biopsies. *Research and Reports in Urology*, 13:799, 2021.

[305] EJ Bass, A Pantovic, M Connor, R Gabe, AR Padhani, A Rockall, H Sokhi, H Tam, M Winkler, and HU Ahmed. A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate mri for prostate cancer in men at risk. *Prostate Cancer and Prostatic Diseases*, 24(3):596–611, 2021.

[306] Yiqiang Zhan and Dinggang Shen. Deformable segmentation of 3-d ultrasound prostate images using statistical texture matching method. *IEEE Transactions on Medical Imaging*, 25(3):256–272, 2006.

[307] Raman Preet Singh, Savita Gupta, and U. Rajendra Acharya. Segmentation of prostate contours for automated diagnosis using ultrasound images: A survey. *Journal of Computational Science*, 21:223–231, 2017. ISSN 1877-7503. doi: https://doi.org/10.1016/j.jocs.2017.04.016. URL https://www.sciencedirect.com/science/article/pii/S1877750317304611.

[308] Sydney Jones and Kevin R Carter. Sonography endorectal prostate assessment, protocols, and interpretation. 2021.

[309] Gaurav Garg and Mamta Juneja. A survey of prostate segmentation techniques in different imaging modalities. *Current Medical Imaging*, 14(1):19–46, 2018.

[310] Ulf-Håkan Stenman, Jari Leinonen, Wan-Ming Zhang, and Patrik Finne. Prostate-specific antigen. *Seminars in Cancer Biology*, 9(2):83–93, 1999. ISSN 1044-579X. doi: https://doi.org/10.1006/scbi.1998.0086. URL https://www.sciencedirect.com/science/article/pii/S1044579X98900864.

[311] Tristan Barrett, Arumugam Rajesh, Andrew B Rosenkrantz, Peter L Choyke, and Baris Turkbey. Pi-rads version 2.1: one small step for prostate mri. *Clinical radiology*, 74(11):841–852, 2019.

[312] Leonard Marks, Shelena Young, and Shyam Natarajan. Mri–ultrasound fusion for guidance of targeted prostate biopsy. *Current opinion in urology*, 23(1):43, 2013.

[313] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[314] Yi Wang, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni. Deep attentional features for prostate segmentation in ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–530. Springer, 2018.

[315] S Sara Mahdavi, Nick Chng, Ingrid Spadinger, William J Morris, and Septimiu E Salcudean. Semi-automatic segmentation for prostate interventions. *Medical Image Analysis*, 15(2):226–237, 2011.

[316] Lixin Gong, Sayan D Pathak, David R Haynor, Paul S Cho, and Yongmin Kim. Parametric shape modeling using deformable superellipses for prostate segmentation. *IEEE transactions on medical imaging*, 23(3):340–349, 2004.

[317] Laurent Saroul, Olivier Bernard, Didier Vray, and Denis Friboulet. Prostate segmentation in echographic images: A variational approach using deformable super-ellipse and rayleigh distribution. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 129–132. IEEE, 2008.

[318] Nicola Altini, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, et al. Liver, kidney and spleen segmentation from ct scans and mri with deep learning: A survey. *Neurocomputing*, 490:30–53, 2022.

[319] Antonio Brunetti, Nicola Altini, Domenico Buongiorno, Emilio Garolla, Fabio Corallo, Matteo Gravina, Vitoantonio Bevilacqua, and Berardino Prencipe. A machine learning and radiomics approach in lung cancer for predicting histological subtype. *Applied Sciences*, 12(12):5829, 2022.

[320] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Francescomaria Marino, Maria Teresa Rocchetti, Silvia Matino, Umberto Venere, Michele Rossini, Francesco Pesce, Loreto Gesualdo, et al., and Vitoantonio Bevilacqua. Semantic segmentation framework for glomeruli detection and classification in kidney histological sections. *Electronics*, 9(3):503, 3 2020. ISSN 2079-9292. doi: 10.3390/electronics9030503.

[321] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, De Irio De Feudis, Domenico Buongiorno, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies. *Electronics*, 9(11):1768, 10 2020. ISSN 2079-9292. doi: 10.3390/electronics9111768.

[322] Liangliang Liu, Jianhong Cheng, Quan Quan, Fang-Xiang Wu, Yu-Ping Wang, and Jianxin Wang. A survey on u-shaped networks in medical image segmentations. *Neurocomputing*, 409:244–258, 2020.

[323] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, 39(12):2481–2495, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2644615.

[324] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28.

[325] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:424–432, 6 2016. ISSN 16113349. doi: 10.1007/978-3-319-46723-8_49.

[326] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. pages 565–571. IEEE, 10 2016. ISBN 978-1-5090-5407-7. doi: 10.1109/3DV.2016.79.

[327] Nicola Altini, Berardino Prencipe, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane, and Giacomo Donato Cascarano. A tversky loss-based convolutional neural network for liver vessels segmentation. pages 342–354. 2020. doi: 10.1007/978-3-030-60799-9_30.

[328] Berardino Prencipe, Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Andrea Guerriero, and Vitoantonio Bevilacqua. Focal dice loss-based v-net for liver segments classification. *Applied Sciences*, 12(7), 2022. ISSN 2076-3417. doi: 10.3390/app12073247. URL https://www.mdpi.com/2076-3417/12/7/3247.

[329] Vitoantonio Bevilacqua, Nicola Altini, Berardino Prencipe, Antonio Brunetti, Laura Villani, Antonello Sacco, Chiara Morelli, Michele Ciaccia, and Arnaldo Scardapane. Lung segmentation and characterization in covid-19 patients for assessing pulmonary thromboembolism: An approach based on deep learning and radiomics. *Electronics*, 10 (20):2475, 2021.

[330] Nicola Altini, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibilano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, Ilaria Sabina Tatò, Gioacchino Brunetti, Vito Triggiani, and Vitoantonio Bevilacqua. Segmentation and identification of vertebrae in ct scans using cnn, k-means clustering and k-nn. *Informatics*, 8(2), 2021. ISSN 2227-9709. doi: 10.3390/informatics8020040. URL https://www.mdpi.com/2227-9709/8/2/40.

[331] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *Informatik aktuell*, page 22, 2019. ISSN 1431472X. doi: 10.1007/978-3-658-25326-4_7.

[332] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

[333] Tim McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2):91–108, 1996.

[334] Johan Montagnat, Hervé Delingette, and Nicholas Ayache. A review of deformable surfaces: topology, geometry and deformation. *Image and vision computing*, 19(14): 1023–1040, 2001.

[335] Isaac Bankman. *Handbook of medical image processing and analysis*. Elsevier, 2008.

[336] Paul J Besl. Geometric modeling and computer vision. *Proceedings of the IEEE*, 76 (8):936–958, 1988.

[337] Richard J Campbell and Patrick J Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.

[338] Ismail B Tutar, Sayan D Pathak, Lixin Gong, Paul S Cho, Kent Wallner, and Yongmin Kim. Semiautomatic 3-d prostate segmentation from trus images using spherical harmonics. *IEEE transactions on medical imaging*, 25(12):1645–1654, 2006.

[339] Michael Unser, Akram Aldroubi, and Murray Eden. B-spline signal processing. i. theory. *IEEE transactions on signal processing*, 41(2):821–833, 1993.

[340] Alan H Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23, 1981.

[341] Alex P Pentland. Perceptual organization and the representation of natural form. In *Readings in Computer Vision*, pages 680–699. Elsevier, 1987.

[342] Alan H Barr. Global and local deformations of solid primitives. *Readings in Computer Vision*, 1(1):661–670, 1987.

[343] Franc Solina and Ruzena Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE transactions on pattern analysis and machine intelligence*, 12(2):131–147, 1990.

[344] S. Pieper, M. Halle, and R. Kikinis. 3d slicer. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 632–635 Vol. 1, 2004. doi: 10.1109/ISBI.2004.1398617.

[345] Andriy Fedorov, Siavash Khallaghi, C Antonio Sánchez, Andras Lasso, Sidney Fels, Kemal Tuncali, Emily Neubauer Sugar, Tina Kapur, Chenxi Zhang, William Wells, et al. Open-source image registration for mri–trus fusion-guided prostate interventions. *International journal of computer assisted radiology and surgery*, 10(6):925–934, 2015.

[346] Calvin R Maurer, Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, 2003.

[347] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Josa a*, 4(4):629–642, 1987.

[348] Daniel N Costa, Ivan Pedrosa, Francisco Donato Jr, Claus G Roehrborn, and Neil M Rofsky. Mr imaging–transrectal us fusion for targeted prostate biopsies: implications for diagnosis and clinical management. *Radiographics*, 35(3):696–708, 2015.

[349] Nicola Altini, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane, and Vitoantonio Bevilacqua. Liver, kidney and spleen segmentation from ct scans and mri with deep learning: A survey. *Neurocomputing*, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.08.157. URL https://www.sciencedirect.com/science/article/pii/S0925231222003149.

[350] Young Jae Kim, Bilegt Ganbold, and Kwang Gi Kim. Web-based spine segmentation using deep learning in computed tomography images. *Healthcare Informatics Research*, 26(1):61–67, 2020. ISSN 2093369X. doi: 10.4258/hir.2020.26.1.61.

[351] Malinda Vania, Dawit Mureja, and Deukhee Lee. Automatic spine segmentation from ct images using convolutional neural network via redundant generation of class labels. *Journal of Computational Design and Engineering*, 6(2):224–232, 2019. ISSN 22885048. doi: 10.1016/j.jcde.2018.05.002.

[352] Syed Furqan Qadri, Danni Ai, Guoyu Hu, Mubashir Ahmad, Yong Huang, Yongtian Wang, and Jian Yang. Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in ct images. *Applied Sciences*, 9(1), 2019. ISSN 2076-3417. doi: 10.3390/app9010069. URL https://www.mdpi.com/2076-3417/9/1/69.

[353] Nikolas Lessmann, Bram van Ginneken, Pim A. de Jong, and Ivana Išgum. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, 53:142–155, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.02.005.

[354] Mina Zareie, Hossein Parsaei, Saba Amiri, · Malik, Shahzad Awan, and Mohsen Ghofrani. Automatic segmentation of vertebrae in 3D CT images using adaptive fast 3D pulse coupled neural networks. *Australasian Physical & Engineering Sciences in Medicine*, 41:1009–1020, 2018. doi: 10.1007/s13246-018-0702-3. URL https://doi.org/10.1007/s13246-018-0702-3.

[355] Anjany Sekuboyina, Amirhossein Bayat, Malek E. Husseini, Maximilian Löffler, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang,

Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olaguer, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Qingyue Wei, Kevin Brown, Matthias Wolf, Alexandre Kirszenberg, Élodie Puybareauq, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. 2020.

[356] Alexandra L Williams, Aisha Al-Busaidi, Patrick J Sparrow, Judith E Adams, and Richard W Whitehouse. Under-reporting of osteoporotic vertebral fractures on computed tomography. *European journal of radiology*, 69(1):179–183, 1 2009. ISSN 1872-7727 (Electronic). doi: 10.1016/j.ejrad.2007.08.028.

[357] R Korez, B Ibragimov, B Likar, F Pernuš, and T Vrtovec. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. *IEEE Transactions on Medical Imaging*, 34(8):1649–1662, 2015. ISSN 1558-254X. doi: 10.1109/TMI.2015.2389334.

[358] Jianhua Yao, Joseph E Burns, Daniel Forsberg, Alexander Seitel, Abtin Rasoulian, Purang Abolmaesumi, Kerstin Hammernik, Martin Urschler, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Isaac Castro-Mateos, Jose M Pozo, Alejandro F Frangi, Ronald M Summers, and Shuo Li. A multi-center milestone study of clinical vertebral ct segmentation. *Computerized Medical Imaging and Graphics*, 49:16–28, 2016. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2015.12.006.

[359] Maximilian T. Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190138.

[360] Anjany Sekuboyina, Markus Rempfler, Alexander Valentinitsch, Bjoern H. Menze, and Jan S. Kirschke. Labeling vertebrae with two-dimensional reformations of multidetector ct images: An adversarial approach for incorporating prior knowledge of spine anatomy. *Radiology: Artificial Intelligence*, 2(2):e190074, 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190074.

[361] Jianhua Yao, Joseph E Burns, Hector Munoz, and Ronald M Summers. Detection of vertebral body fractures based on cortical shell unwrapping. pages 509–516, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33454-2.

[362] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[363] Qian Chang, Jun Shi, and Zhiheng Xiao. A new 3d segmentation algorithm based on 3d pcnn for lung ct slices. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–5, 2009. doi: 10.1109/BMEI.2009.5305554.

[364] Hyun Jin Bae, Heejung Hyun, Younghwa Byeon, Keewon Shin, Yongwon Cho, Young Ji Song, Seong Yi, Sung Uk Kuh, Jin S. Yeom, and Namkug Kim. Fully automated 3d segmentation and separation of multiple cervical vertebrae in ct images using a 2d convolutional neural network. *Computer Methods and Programs in Biomedicine*, 184, 2020. ISSN 18727565. doi: 10.1016/j.cmpb.2019.105119.

[365] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 5(Visigrapp):124–133, 2020. doi: 10.5220/0008975201240133.

[366] Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitsch, Jan S. Kirschke, and Bjoern H. Menze. Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11073 LNCS(October):649–657, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00937-3_74.

[367] Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. pages 262–270, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40763-5.

[368] Ben Glocker, J Feulner, Antonio Criminisi, D R Haynor, and E Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. pages 590–598, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33454-2.

[369] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 5 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

[370] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. 11 2014.

[371] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019. ISSN 18728286. doi: 10.1016/j.neucom.2019.02.003.

[372] Judith E. Adams, Zulf Mughal, John Damilakis, and Amaka C. Offiah. *Radiology*. 2012. ISBN 9780123820402. doi: 10.1016/B978-0-12-382040-2.10012-7.

[373] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[374] Chen Shen, Fausto Milletari, Holger R. Roth, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. Improving v-nets for multi-class abdominal organ segmentation. page 10. SPIE, 3 2019. ISBN 9781510625457. doi: 10.1117/12.2512790.

[375] Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. 3 2020.

[376] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[377] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk. Nearest-neighbor methods in learning and vision. In *Neural Information Processing*, 2005.

[378] Matthew McCormick, Xiaoxiao Liu, Luis Ibanez, Julien Jomier, and Charles Marion. Itk: enabling reproducible research and open science. *Frontiers in Neuroinformatics*, 8: 13, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00013.

[379] Schroeder Will; Martin Ken; Lorensen Bill. *The Visualization Toolkit (4th ed.)*. 2006.

[380] G Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

[381] Tobias Heimann, Brain van Ginneken, M.A. Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M.M. Cashman, Ying Chi, Andrés Córdova, B.M. Benoit M. Dawant, Márta Fidrich, J.D. Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar Kainmüller, R.I. Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, H.-P. Hans Peter Meinzer, Gábor Németh, D.S. Daniela S. Raicu, A.-M. Anne Mareike Rau, Eva M. E.M. van Rikxoort, Mikaël Rousson, Lászlo Ruskó, K.A. Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. J.M. Waite, Andreas Wimmer, Ivo Wolf, Ying Chi, A. Cordova, B.M. Benoit M. Dawant, Márta Fidrich, J.D. Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, D. Kainmuller, R.I. Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, H.-P. Hans Peter Meinzer, G. Nemeth, D.S. Daniela S. Raicu, A.-M. Anne Mareike Rau, Eva M. E.M. van Rikxoort, Mikaël Rousson, L. Rusko, K.A. Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. J.M. Waite, Andreas Wimmer, and Ivo Wolf. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 8 2009. ISSN 0278-0062. doi: 10.1109/TMI.2009.2013851.

[382] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, Samuel Kadoury, Tomasz Konopczynski, Miao Le, Chunming Li, Xiaomeng Li, Jana Lipkovà, John Lowengrub, Hans Meine, Jan Hendrik Moltz, Chris Pal, Marie Piraud, Xiaojuan Qi, Jin Qi, Markus Rempfler, Karsten Roth, Andrea Schenk, Anjany Sekuboyina, Eugene Vorontsov, Ping Zhou, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Felix Gruen, Georgios Kaissis, Fabian Lohöfer, Rickmer Braren, Julian Holch, Felix Hofmann, Wieland Sommer, Volker Heinemann, Colin Jacobs, Gabriel Efrain Humpire Mamani, Bram van Ginneken, Gabriel Chartrand, An Tang, Michal Drozdzal, Avi Ben-Cohen, Eyal Klang, Marianne M. Amitai, Eli Konen, Hayit Greenspan, Johan Moreau, Alexandre Hostettler, Luc Soler, Refael Vivanti, Adi Szeskin, Naama Lev-Cohain, Jacob Sosna, Leo Joskowicz, and Bjoern H. Menze. The liver tumor segmentation benchmark (lits). pages 1–43, 1 2019.

[383] Berardino Prencipe, Nicola Altini, Giacomo Donato Cascarano, Andrea Guerriero, and Antonio Brunetti. A novel approach based on region growing algorithm for liver and spleen segmentation from ct scans. In De-Shuang Huang, Vitoantonio Bevilacqua, and Abir Hussain, editors, *Intelligent Computing Theories and Application*, pages 398–410, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60799-9.

[384] Vitoantonio Bevilacqua, Antonio Brunetti, Andrea Guerriero, Gianpaolo Francesco Trotta, Michele Telegrafo, and Marco Moschetta. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cognitive Systems Research*, 53:3–19, 2019.

[385] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahttps Jemal. Cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 68(6):394–424, 2018.

[386] Marzieh Esmaeili, Seyed Mohammad Ayyoubzadeh, Zohreh Javanmard, and Sharareh R Niakan Kalhori. A systematic review of decision aids for mammography screening: Focus on outcomes and characteristics. *International Journal of Medical Informatics*, 149:104406, 2021.

[387] Supriya Kulkarni, Vivianne Freitas, and Derek Muradali. Digital breast tomosynthesis: potential benefits in routine clinical practice. *Canadian Association of Radiologists Journal*, page 08465371211025229, 2021.

[388] Mingxiang Wu and Jie Ma. Association between imaging characteristics and different molecular subtypes of breast cancer. *Academic Radiology*, 24(4):426–434, 2017.

[389] Si-Qing Cai, Jian-Xiang Yan, Qing-Shi Chen, Mei-Ling Huang, and Dong-Lu Cai. Significance and application of digital breast tomosynthesis for the bi-rads classification of breast cancer. *Asian Pacific Journal of Cancer Prevention*, 16(9):4109–4114, 2015.

[390] E Sickles, CJ D'Orsi, and LW Bassett. Acr bi-rads® mammography. acr bi-rads® atlas, breast imaging reporting and data system. american college of radiology 2013.

[391] Su Hyun Lee, Jung Min Chang, Sung Ui Shin, A Jung Chu, Ann Yi, Nariya Cho, and Woo Kyung Moon. Imaging features of breast cancers on digital breast tomosynthesis according to molecular subtype: association with breast cancer detection. *The British Journal of Radiology*, 90(1080):20170470, 2017.

[392] Siqing Cai, Miaomiao Yao, Donglu Cai, Jianxiang Yan, Meiling Huang, Lisheng Yan, and Huirong Huang. Association between digital breast tomosynthesis and molecular subtypes of breast cancer. *Oncology Letters*, 17(3):2669–2676, 2019.

[393] Vitoantonio Bevilacqua. Three-dimensional virtual colonoscopy for automatic polyps detection by artificial neural network approach: New tests on an enlarged cohort of polyps. *Neurocomputing*, 116:62–75, 2013.

[394] Vitoantonio Bevilacqua, Antonio Brunetti, Gianpaolo Francesco Trotta, Giovanni Dimauro, Katarina Elez, Vito Alberotanza, and Arnaldo Scardapane. A novel approach for hepatocellular carcinoma detection and classification based on triphasic ct protocol. In *2017 IEEE congress on evolutionary computation (CEC)*, pages 1856–1863. IEEE, 2017.

[395] Jiande Wu and Chindo Hicks. Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2):61, 2021.

[396] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6, 2019.

[397] Samir S Yadav and Shivajirao M Jadhav. Thermal infrared imaging based breast cancer diagnosis using machine learning techniques. *Multimedia Tools and Applications*, pages 1–19, 2020.

[398] Dina A Ragab, Omneya Attallah, Maha Sharkas, Jinchang Ren, and Stephen Marshall. A framework for breast cancer classification using multi-dcnns. *Computers in Biology and Medicine*, 131:104245, 2021.

[399] Mohammad M Ghiasi and Sohrab Zendehboudi. Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, 128:104089, 2021.

[400] Yu-Dong Zhang, Suresh Chandra Satapathy, David S Guttery, Juan Manuel Górriz, and Shui-Hua Wang. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing & Management*, 58(2):102439, 2021.

[401] Raouia Mokni, Norhene Gargouri, Alima Damak, Dorra Sellami, Wiem Feki, and Zeineb Mnif. An automatic computer-aided diagnosis system based on the multimodal fusion of breast cancer (mf-cad). *Biomedical Signal Processing and Control*, 69:102914, 2021.

[402] Jiaqiao Shi, Aleksandar Vakanski, Min Xian, Jianrui Ding, and Chunping Ning. Emt-net: Efficient multitask network for computer-aided diagnosis of breast cancer. *arXiv preprint arXiv:2201.04795*, 2022.

[403] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Gene Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908, 2021.

[404] Nasibeh Saffari, Hatem A Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Meritxell Arenas, Eleni Mangina, Blas Herrera, and Domenec Puig. Fully automated breast density segmentation and classification using deep learning. *Diagnostics*, 10(11): 988, 2020.

[405] Neeraj Shrivastava and Jyoti Bharti. Breast tumor detection and classification based on density. *Multimedia Tools and Applications*, 79(35):26467–26487, 2020.

[406] Vivek Kumar Singh, Hatem A Rashwan, Santiago Romani, Farhan Akram, Nidhi Pandey, Md Mostafa Kamal Sarker, Adel Saleh, Meritxell Arenas, Miguel Arquez, Domenec Puig, et al. Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications*, 139:112855, 2020.

[407] Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro. Icadx: interpretable computer aided diagnosis of breast masses. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 1057522. International Society for Optics and Photonics, 2018.

[408] Per Skaane, Andriy I Bandos, Loren T Niklason, Sofie Sebuødegård, Bjørn H Østerås, Randi Gullien, David Gur, and Solveig Hofvind. Digital mammography versus digital mammography plus tomosynthesis in breast cancer screening: the oslo tomosynthesis screening trial. *Radiology*, 291(1):23–30, 2019.

[409] Xin Li, Genggeng Qin, Qiang He, Lei Sun, Hui Zeng, Zilong He, Weiguo Chen, Xin Zhen, and Linghong Zhou. Digital breast tomosynthesis versus digital mammography: integration of image modalities enhances deep learning-based breast mass classification. *European radiology*, 30(2):778–788, 2020.

[410] Kayla Mendel, Hui Li, Deepa Sheth, and Maryellen Giger. Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. *Academic radiology*, 26(6): 735–743, 2019.

[411] Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine & Biology*, 63(9):095005, 2018.

[412] Sergei V Fotin, Yin Yin, Hrishikesh Haldankar, Jeffrey W Hoffmeister, and Senthil Periaswamy. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97850X. International Society for Optics and Photonics, 2016.

[413] SKM Hamouda, RHB El-Ezz, and Mohammed E Wahed. Enhancement accuracy of breast tumor diagnosis in digital mammograms. *Journal of Biomedical Sciences*, 6(4): 1–8, 2017.

[414] Ayaka Sakai, Yuya Onishi, Misaki Matsui, Hidetoshi Adachi, Atsushi Teramoto, Kuniaki Saito, and Hiroshi Fujita. A method for the automated classification of benign and malignant masses on digital breast tomosynthesis images using machine learning and radiomic features. *Radiological Physics and Technology*, 13(1):27–36, 2020.

[415] Said Boumaraf, Xiabi Liu, Chokri Ferkous, and Xiaohong Ma. A new computer-aided diagnosis system with modified genetic feature selection for bi-rads classification of breast masses in mammograms. *BioMed Research International*, 2020, 2020.

[416] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[417] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[418] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[419] Guang Yang, Qinghao Ye, and Jun Xia. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77:29–52, 2022.

[420] R Ricciardi, G Mettivier, M Staffa, A Sarno, G Acampora, S Minelli, A Santoro, E Antignani, A Orientale, IAM Pilotti, et al. A deep learning classifier for digital breast tomosynthesis. *Physica Medica*, 83:184–193, 2021.

[421] Mehedi Masud, Amr E Eldin Rashed, and M Shamim Hossain. Convolutional neural network-based models for diagnosis of breast cancer. *Neural Computing and Applications*, pages 1–12, 2020.

[422] Yong Joon Suh, Jaewon Jung, and Bum-Joo Cho. Automated breast cancer detection in digital mammograms of various densities via deep learning. *Journal of personalized medicine*, 10(4):211, 2020.

[423] Meng Lou, Runze Wang, Yunliang Qi, Wenwei Zhao, Chunbo Xu, Jie Meng, Xiangyu Deng, and Yide Ma. Mgbn: Convolutional neural networks for automated benign and malignant breast masses classification. *Multimedia Tools and Applications*, 80(17): 26731–26750, 2021.

[424] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[425] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[426] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[427] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[428] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[429] Hui Wen Loh, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, page 107161, 2022.

[430] Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, and Natalia Díaz-Rodríguez. Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv:1909.09065*, 2019.

[431] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[432] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer, 2022.

[433] Christine T Wolf and Kathryn E Ringland. Designing accessible, explainable ai (xai) experiences. *ACM SIGACCESS Accessibility and Computing*, (125):1–1, 2020.

[434] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

[435] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[436] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[437] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[438] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[439] Stefan Buijsman. Defining explanation and explanatory depth in xai. *Minds and Machines*, 32(3):563–584, 2022.

[440] Robert R Hoffman, Gary Klein, and Shane T Mueller. Explaining explanation for “explainable ai”. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[441] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

[442] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[443] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[444] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[445] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John-Dylan Haynes, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation. *NeuroImage: Clinical*, 24:102003, 2019.

[446] Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: from black box to interpretable models. In *Embedded Systems and Artificial Intelligence*, pages 327–337. Springer, 2020.

[447] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[448] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, 114:107856, 2021.

[449] Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022.

[450] Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

[451] Cristiano Patrício, João C Neves, and Luís F Teixeira. Explainable deep learning methods in medical diagnosis: A survey. *arXiv preprint arXiv:2205.04766*, 2022.

[452] Miquel Miró-Nicolau, Gabriel Moyà-Alcover, and Antoni Jaume-i Capó. Evaluating explainable artificial intelligence for x-ray image analysis. *Applied Sciences*, 12(9):4459, 2022.

[453] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[454] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[455] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[456] Abhaya Agarwal and Alon Lavie. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*, 2007.

[457] Florian Dubost, Hieab Adams, Pinar Yilmaz, Gerda Bortsova, Gijs van Tulder, M Arfan Ikram, Wiro Niessen, Meike W Vernooij, and Marleen de Bruijne. Weakly supervised object detection with 2d and 3d regression neural networks. *Medical Image Analysis*, 65: 101767, 2020.

[458] Sema Candemir, Richard D White, Mutlu Demirer, Vikash Gupta, Matthew T Bigelow, Luciano M Prevedello, and Barbaros S Erdal. Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary ct angiography with a deep 3-dimensional convolutional neural network. *Computerized Medical Imaging and Graphics*, 83:101721, 2020.

[459] Claire Tang. Discovering unknown diseases with explainable automated medical imaging. In *Annual Conference on Medical Image Understanding and Analysis*, pages 346–358. Springer, 2020.

[460] Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE transactions on medical imaging*, 39(2):413–424, 2019.

[461] Chao Cong, Yoko Kato, Henrique Doria Vasconcellos, Joao Lima, and Bharath Venkatesh. Automated stenosis detection and classification in x-ray angiography using deep neural network. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1301–1308. IEEE, 2019.

[462] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. *Frontiers in aging neuroscience*, 11:194, 2019.

[463] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 69–75, 2017.

[464] Dasom Seo, Kanghan Oh, and Il-Seok Oh. Regional multi-scale approach for visually pleasing explanations of deep neural networks. *IEEE Access*, 8:8572–8582, 2019.

[465] Baihong Xie, Ting Lei, Nan Wang, Hongmin Cai, Jianbo Xian, Miao He, Lihe Zhang, and Hongning Xie. Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 15(8):1303–1312, 2020.

[466] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1):1–12, 2020.

[467] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8309–8319, 2018.

[468] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387*, 2019.

[469] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[470] Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. From chest x-rays to radiology reports: a multimodal machine learning approach. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2019.

[471] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673, 2019.

[472] Graham Spinks and Marie-Francine Moens. Justifying diagnosis decisions by deep neural networks. *Journal of biomedical informatics*, 96:103248, 2019.

[473] Ivan Rodin, Irina Fedulova, Artem Shelmanov, and Dmitry V Dylov. Multitask and multimodal neural network model for interpretable analysis of x-ray images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1601–1604. IEEE, 2019.

[474] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.