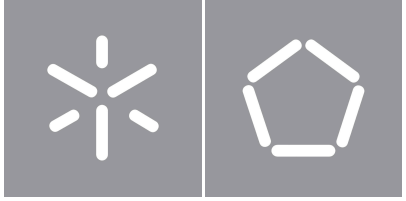
University of Minho
School of Engineering

Eduardo Jorge Santos Teixeira

Deep Learning approaches for Landmarks Detection on Knee Medical Images

july 2024



University of Minho
School of Engineering

Eduardo Jorge Santos Teixeira

Deep Learning approaches for Landmarks Detection on Knee Medical Images

Master's in Informatics Engineering

Deep Learning Specialization
Work conducted under the guidance of
Professor Doctor Cristina P. Santos

july 2024

Copyright and Terms of Use for Third Party Work

License granted to users of this work:

Acknowledgements

A expressão “O fim” tem muitas vezes uma conotação negativa em diferentes contextos. No entanto, neste caso, não há nada mais satisfatório do que poder declarar “o fim” de um projeto desta dimensão e complexidade. Foram muitos anos e incontáveis horas dedicadas a atingir este marco, e é com imenso orgulho e agradecimento que expresso este sentimento. Gostaria de estender os meus mais profundos agradecimentos a diferentes pessoas, cuja orientação e encorajamento foram fundamentais para a conclusão deste projeto.

Em primeiro lugar, à minha família, obrigado pelo vosso apoio inabalável e pela vossa paciência. À mãe e ao pai, a vossa crença em mim manteve-me motivado e concentrado, mesmo nos momentos mais desafiantes e complicados, por me poderem ajudar, partilhando a falta de sono e o stress. À minha irmã Filipa, que compreendeu desde o início a árdua tarefa que me propus, ajudando com as suas receitas mágicas e a capacidade de arranjar um companheiro informático, o Adriano, que me forneceu, sem hesitação, um computador crucial ao projeto. À minha irmã Sarah, que todos os dias me faz refletir sobre a vivência e a simplicidade do presente.

Estou grato à Professora Cristina P. Santos e à equipa do BirdLab, especificamente à Diana e ao Roberto, pela sua inestimável orientação, experiência e apoio ao longo desta jornada. Os vossos conhecimentos e encorajamento foram cruciais para a conclusão bem-sucedida deste trabalho, em particular, pela paciência em suportar a minha busca incessante da perfeição e obsessão pelos joelhos.

Um sincero agradecimento aos meus colegas, amigos e companheiros de casa, em particular ao Carlos, Gonçalo, Soares e Sara. Aqueles momentos de happy hour fizeram-me perceber como este percurso teve momentos agradáveis e enriquecedores. A vossa simples presença foi fundamental para o meu sucesso. Por fim, gostaria de agradecer a todos os que me apoiaram, direta ou indiretamente, na realização deste projeto. Os vossos contributos foram profundamente apreciados. A conclusão deste projeto foi um percurso notável e estou profundamente grato pelo apoio que recebi ao longo do caminho.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, july 2024

Eduardo Jorge Santos Teixeira

Resumo

A Instabilidade Patelofemoral (PFI) é uma das condições mais comuns que afetam o joelho, especialmente em jovens entre 10 e 17 anos (incidência 5,8-29 por 100.000 pessoas), com maior incidência no sexo feminino. Fatores de risco incluem a forma da tróclea, o posicionamento da patela, a distância excessiva entre a tuberosidade tibial e a tróclea, eventos traumáticos, e a frouxidão ou relaxamento dos ligamentos, resultando em problemas como rutura do ligamento medial patelofemoral e dor crónica no joelho. A avaliação de PFI enfrenta variabilidade e falta de padronização varia o processo entre hospitais, levando ao uso de diferentes combinações de índices, exigindo uma ferramenta que simplifique e padronize o processo.

Esta dissertação visa desenvolver e otimizar modelos de Aprendizagem Profunda, incluindo Redes neurais Convolucionais (CNN), para detetar pontos de referência em imagens 3D do joelho, com foco na PFI. Para isso foi desenvolvido um *pipeline* baseado na metodologia CRISP-DM, desde a análise de ficheiros DICOM até à preparação dos dados 3D. A estrutura incluiu a constituição de 3 subsets rotulados (Axial, Sagittal e Dynamic) e a preparação dos dados para modelagem (análise descritiva, reamostragem e construção de *ground truth*). Foi então construída uma interface gráfica de utilizador (GUI) para marcar pontos anatómicos com precisão, garantindo uma rotulagem abrangente dos índices de PFI.

Modelos baseados em CNN de segmentação 3D U-Net com ligações residuais e mecanismos de atenção foram treinados e avaliados. Blocos de *downsampling* e *upsampling* foram construídos com convoluções padrão e transpostas, incluindo conexões residuais para facilitar o fluxo de gradientes. Técnicas de *data augmentation* e ajustes de hiperparâmetros, como taxa de aprendizagem, normalização de gradiente, funções de perda e de ativação, foram aplicados para melhorar a robustez e o desempenho dos modelos. Diferentes *callbacks* foram utilizados para monitorar o treino e melhorar a performance do modelo.

Os resultados no subset Axial, usando validação *5-fold*, foram abaixo do esperado, com Mean Absolute Error (MAE) de 42.34 mm. No subset Sagittal, a validação *5-fold* mostrou consistência elevada com MAE médio de 1.65 mm. Ambos usaram o modelo 3D U-Net Simples. No subset Dynamic, o melhor desempenho foi do modelo 3D U-Net com mecanismos residuais e de atenção, com MAE médio de 3.69 mm. A avaliação do subset Dynamic não utilizou validação *5-fold*, destacando a necessidade desta para resultados mais confiáveis.

Em conclusão, as abordagens de Aprendizagem Profunda propostas são eficazes para a deteção de pontos de referência em imagens do joelho, e em específico para a otimização do diagnóstico da PFI. Estes métodos podem ser integrados em sistemas de diagnóstico assistido por computador, contribuindo para diagnósticos mais precisos, automatizados e intervenções cirúrgicas mais seguras. Futuras pesquisas devem explorar não só diferentes e emergentes técnicas de Aprendizagem Profunda, como a melhoria contínua dos modelos ao nível de *fine tuning*.

Palavras-chave Aprendizagem Profunda, Imagem médica, Instabilidade patelofemoral, Joelho, Deteção de pontos de referência, Redes Neuronais Convolucionais

Abstract

Patellofemoral Instability (PFI) is one of the most common conditions affecting the knee, especially in young people aged 10 to 17 years (incidence 5.8-29 per 100,000 people), with a higher incidence in females. Risk factors include trochlear shape, patella positioning, excessive tibial tuberosity-trochlear distance, traumatic events, and ligament laxity or relaxation, resulting in issues such as medial patellofemoral ligament rupture and chronic knee pain. PFI assessment faces variability and lack of standardisation, varying the process between hospitals, leading to the use of different combinations of indices, thus requiring a tool that simplifies and standardises the process.

This dissertation aims to develop and optimise Deep Learning models, including convolutional neural networks (CNN), to detect reference points in 3D knee images, focusing on PFI. A pipeline based on the CRISP-DM framework was developed, from DICOM file analysis to 3D data preparation. The structure included the creation of three labelled subsets (Axial, Sagittal, and Dynamic) and data preparation for modelling (descriptive analysis, resampling, and ground truth construction). A graphical user interface (GUI) was built to accurately mark anatomical points, ensuring comprehensive labelling of PFI indices.

3D U-Net segmentation CNN models with residual connections and attention mechanisms were trained and evaluated. Downsampling and upsampling blocks were constructed with standard and transposed convolutions, including residual connections to facilitate gradient flow. Data augmentation techniques and hyperparameter adjustments, such as learning rate, gradient normalisation, loss and activation functions, were applied to improve model robustness and performance. Various callbacks were used to monitor training and enhance model performance.

The results for the Axial subset, using 5-fold validation, were below expectations, with an MAE of 42.34 mm. In the Sagittal subset, 5-fold validation showed high consistency with an average MAE of 1.65 mm. Both used the simple 3D U-Net model. In the Dynamic subset, the best performance was achieved by the 3D U-Net model with residual connections and attention mechanisms, with an average MAE of 3.69 mm. The Dynamic subset evaluation did not use 5-fold validation, highlighting the need for this for more reliable results.

In conclusion, the proposed Deep Learning approaches are effective for detecting reference points in knee images, specifically optimising PFI diagnosis. These methods can be integrated into computer-assisted diagnostic systems, contributing to more accurate, automated diagnoses and safer surgical interventions. Future research should explore not only different and emerging Deep Learning techniques but also the continuous improvement of models through fine-tuning.

Keywords Convolutional Neural Networks, Deep Learning, Knee, Landmark Detection, Medical Imaging, Patellofemoral Instability

Contents

List of Figures

List of Tables

List of Equations

Acronyms

2D Two Dimensional.

3D Three Dimensional.

AI Artificial Intelligence.

BiRDLab Biomedical Robotic Devices Laboratory.

BU-Net Bayesian U-Net.

CAD Computer-Assisted Diagnosis.

CBCT Cone Beam Computer Tomography.

CMEMS Center of MicroElectroMechanical Systems.

CoM Center of Mass.

CT Computed Tomography.

DICOM Digital Imaging and Communications in Medicine.

DL Deep Learning.

FCNN Fully Convolutional Neural Network.

FPN Feature Pyramid Networks.

GANs Generative Adversarial Networks.

GPUs Graphics Processing Units.

GUI Graphical User Interface.

JSON JavaScript Object Notation.

Leaky ReLU Leaky Rectified Linear Unit.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MIA Medical Image Analysis.

ML Machine Learning.

MLPs Multi-Layer Perceptrons.

MRA Cerebral Magnetic Resonance Angiography.

MRI Magnetic Resonance Image.

MSE Mean Squared Error.

OCT Optical Coherence Tomography.

OOM Out-Of-Memory.

PD Proton Density.

PFI Patellofemoral Instability.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analysis.

ReLU Rectified Linear Unit.

ResNet Residual Networks.

RMSE Root Mean Squared Error.

RNN Recurrent Neural Networks.

ROI Region of Interest.

RPN Region Proposal Networks.

RQs Research Questions.

SDR Sucess Detection Rate.

TAVI Transcatheter Aortic Valve Implantation.

US Ultrasound.

Chapter 1

Introduction

This dissertation presents the work proposal developed in the scope of the second year of the Master's in Informatics Engineering, at the University of Minho, during the academic years of 2022-2023.

The academic year was passed working in and with the Biomedical Robotic Devices Laboratory (BiRDLab), included in the Center of MicroElectroMechanical Systems (CMEMS) Research Center, at the University of Minho, Guimarães, Portugal. During this time, a tool for the automatic diagnosis of Patellofemoral Instability (PFI) using Deep Learning (DL) algorithms and medical images was developed and validated. This clinical tool was designed to assist radiologists and standardise the diagnostic process, through the accurate detection of anatomical landmarks. All the methods, results, and conclusions are detailed in this document.

1.1 Motivation and Problem Statement

PFI is one of the problems that affects the knee and is characterised by the instability of the patella in relation to the femoral trochlea, which can lead to subluxations or luxations of the patella. It can be caused either by traumatic events of by predisposing anatomical risk factors, leading to an unphysiological sequence of movement within the patellofemoral joint, also known as patellofemoral maltracking [1]. The main potential risk factors leading to PFI are related to trochlear shape, patella positioning and excessive tibial tubercle-trochlear groove distance, resulting in patellofemoral ligament rupture, articular cartilage damage, chronic knee pain [2] and ligament laxity or relaxation [3], impacting patients' locomotion and physical activity [4]. It mainly affects the general population of young people between 10 and 17 years old and females. The incidence of PFI varies between 5.8-29 per 100,000 in the 10 to 17-year-old age group [5].

The diagnosis requires, in addition to the patient's clinical history and the orthopaedic surgeon's physical examination, the acquisition and rigorous analysis of medical images. This includes X-ray, Computed Tomography (CT), and Magnetic Resonance Image (MRI) (an effective and robust modality as it allows the visualisation of articular cartilage, subchondral bones, meniscus and soft tissues) [6]. Radiologists have the difficult task of finding different anatomical landmarks to obtain PFI indexes that are essential for diagnosis [7], [8]. The conventional method of diagnosis is time-consuming and manually performed, leading to significant intra- and inter-observer variability with a lack of precision [9], [10]. Moreover, there is no standard protocol for the diagnostic process, which varies from hospital to hospital according to the preferences of doctors and radiologists. This variability results in different hospitals and clinical centers using a combination of different PFI indexes for diagnosis [11]. The absence of a standardised approach can lead to inconsistent diagnostic outcomes and jeopardises the ability to compare results across different institutions. To address these limitations, there is a critical need for a tool that automates the diagnostic process. Such a tool would not only reduce the time and effort required for diagnosis but also minimize observer variability

and enhance precision. By standardizing the diagnostic procedure, an automated tool would ensure consistent and accurate results, facilitating better clinical decision and diagnosis.

Computer-Assisted Diagnosis (CAD) is therefore becoming increasingly important and necessary. It has been thriving in recent years due to new advances in software, hardware and image quality of the different types of scanners [12]. The application of Artificial Intelligence (AI), particularly Machine Learning (ML) in CAD, offers significant advantages in medical imaging by enabling the automation and standardization of the diagnostic process [13]. ML techniques help CAD systems by analysing vast amounts of medical data to identify patterns and make accurate predictions and classifications, thereby assisting radiologists and clinicians in diagnosing diseases more efficiently.

Within this broad field of ML, DL stands out due to its ability to handle large volumes of data and automatically extract complex features without manual intervention, helping CAD to standardise, normalise and automate the diagnosis process [14]. Under this sub field's diverse technologies are Fast vector , part of this DL progress as they are used for several tasks [15]. CNN are well suited for anatomical landmark detection in medical images as they can learn local patterns in the image, are invariant to small translations, can process images at different scales, and can learn multiple abstraction levels of the input medical image [16]. This optimises and automates the diagnosis, leading to faster, more accurate and objective predictions of PFI indexes, as well as reducing the complexity and duration of the diagnostic task and intra- and inter-observer variability [12], [17].

1.2 Goals and Research Questions

This dissertation aimed to develop DL approaches to detect anatomical landmarks on knee images and automatically diagnose PI. From creating a program to effectively perform labelling to DL-based approaches on constructed data, the overall pipeline solution integrates various stages of data processing, model training, and evaluation. This approach ensures a robust and efficient workflow, ultimately enhancing the accuracy and reliability of PFI diagnosis.

To reach this main goal, it is necessary to achieve the following step-objectives:

Objective 1 : Literature review of the existing DL approaches for landmark detection on medical images. Addressed in Chapter 3.

Objective 2 : Construct a system pipeline architecture for the automatic detection of anatomical landmarks in medical images. Specifying system hardware and software requirements, setting up a robust data processing pipeline, and ensuring the system's compatibility with DL models. Addressed in Chapter 4.

Objective 3 : Design and develop an application to label the data. Labelling process consisting in the detection (position) of anatomical landmarks for the entire dataset landmarks. Addressed in Chapter 5.

Objective 4 : Develop DL-based algorithms, based on the literature review, for automatic PFI diagnosis. Addressed in Chapter 6.

Objective 5 : Validate the DL-based algorithms with a generalized test subset. Addressed in Chapter 7.

The outlined objectives will allow answering the following Research Questions (RQs):

RQ 1 : What factors should be considered for anatomical landmarks detection on medical images based in DL algorithms? The answer is included in Chapter 3.

RQ 2 : Regarding the landmark labeling program, what are the key landmarks necessary for accurate PFI diagnosis, and how does the developed Graphical User Interface (GUI) tool support their precise annotation? The answer is included in Chapter 5.

RQ 3 : What are the essential components and how should an automated pipeline be constructed for accurate and reliable landmark detection in MRI scans used in PFI diagnosis? The answer is included in Chapter 6.

RQ 4 Do DL algorithms actually bring benefits to the diagnostic process and what are the critical factors influencing their performance? The answer is included in Chapter 7.

1.3 Contribution to Knowledge

The main contributions of this dissertation to knowledge are:

- A review of the most recent trends of DL approaches to landmark detection on medical images;
- A GUI, ImageLabelGUI, for landmark labelling, adjusted to the context of data obtained (MRI data format and knee area images only). An important step for volume labelling and ground truth construction;
- Descriptive dataset analysis from clinical to technical metadata, from key image factors to Digital Imaging and Communications in Medicine (DICOM) specifications required;
- Development and training of DL algorithms with current technologies and normalization functions, tailored for knee landmark detection in medical images.
- Test dataset construction in order to evaluate the performance of developed models against existing solutions (Benchmarking), using pre-considered evaluation metrics for model performance comparison.

The work carried out during this dissertation has led to the writing of the following journal papers:

- **E. Teixeira**, D. Rito, R.M. Barbosa, C. P. Santos, "Deep Learning approaches for landmarks detection on medical images: A systematic review" [submitting]

1.4 Dissertation Outline

This manuscript is organised into 8 chapters.

Chapter 2 provides an overview of DL, covering its evolution, core architectures, learning paradigms (supervised, unsupervised, and hybrid), and applications in medical imaging, in particular for knee landmark detection.

Chapter 3 presents the state-of-the-art of DL approaches for landmarks detection in medical images. The specifications of the included studies are presented and discussed, regarding the body area (area and number of landmarks), dataset (modality [Two Dimensional (2D) or Three Dimensional (3D)], size, and dimension), architecture (models, and variants) and evaluation metrics.

Chapter 4 outlines the solution architecture using a specific methodology. It covers task understanding, data preparation, model development, evaluation, and benchmarking. It details hardware and software requirements, test dataset construction, and the system pipeline.

Chapter 5 describes the system development, including data management, test dataset creation, data structure, landmark annotation, and the data labelling tool.

Chapter 6 details the system development, covering data analysis, preprocessing, ground truth construction, resampling, file formats, and modeling approaches with different architectures and training strategies.

Chapter 7 presents and critically discusses the results of the DL algorithms developed for landmark detection on medical images.

Finally, Chapter 8 addresses the conclusions of this dissertation, answering the RQ and appointing future work.

Chapter 2

Theoretical Overview

This chapter provides a foundational overview of DL as a subfield of ML, it forwards its focus on its application in Medical Image Analysis (MIA) development. The journey begins with a discussion of general DL concepts and exploration of learning paradigms. The analysis assess the cores architectures and sets the stage for broader DL applications. Following this, the chapter narrows its focus to MIA and the role of DL in it, highlighting the initial journey, key importance, and inherent challenges of applying DL in this field. This knowledge is finally refined into to the specific task of landmark detection in knee medical images.

2.1 Introduction to Deep Learning

Since its remarkable reappearance in Hinton et al. (2006) [18], DL has revolutionized the field of ML. By enabling multi-layer neural networks capable of recognizing and interpreting complex patterns in vast datasets, it enhanced performance across a wide range of applications. DL models exploit multiple stages of non-linear information processing to learn hierarchical representations of data. In these models, higher-level features are derived from lower-level ones, facilitating a nuanced understanding of data in diverse contexts. This hierarchical feature learning is fundamental to DL's success and impact in various domains. Its contributions to numerous fields highlight the potential of DL in a wide range of applications and use cases.

To understand the broad application spectrum of DL, it is essential to examine it through the lens of different learning paradigms. By exploring supervised, unsupervised, and hybrid learning approaches, an assessment is developed regarding the diversity of methodologies that drive DL's success across various domains.

2.1.1 Supervised, Unsupervised and Hybrid learning

Real-world problems in various applications fields are usually distinguished according to their learning capabilities. These offer extensive combinations to solve supervised, unsupervised, and hybrid problems, with each of these general classes covering a range of DL architectures. These paradigms are well established within the ML community and together form the backbone of many innovative solutions across different domains.

Supervised learning, characterized by the use of labelled datasets, allows models to learn the mapping from inputs to outputs, making it indispensable for tasks that require explicit predictions, such as image classification, object detection or speech recognition. This direct approach to pattern classification uses discriminative architectures designed to model the posterior distributions of classes given observable data, making them highly efficient for training and testing, as well as flexible and suitable for end-to-end learning of complex systems.

Unsupervised learning, on the other hand, thrives on the challenge of discovering hidden patterns or structures in unlabelled

data. Generative architectures, often associated with this learning paradigm, are designed to characterize the high-order correlation properties of visible data for pattern analysis or synthesis. The concept of unsupervised pre-training emerges as a pivotal strategy, particularly for deep networks, facilitating the process of greedily learning the lower network levels, layer-by-layer, from the bottom up, even when training data is scarce.

Hybrid learning architectures represent a confluence of generative and discriminative models aiming to combine the best of both worlds to enhance learning outcomes. This approach typically involves generative pre-training followed by discriminative fine-tuning, providing an empirical solution to common optimization challenges such as poor local optima (i.e. solutions that are better than neighboring solutions within a certain region, but not necessarily the best overall solutions), which is particularly beneficial in contexts where labelled data is limited. The hybrid model's success lies in its ability to exploit the feature hierarchies of the deep architecture, significantly improving performance on a variety of perceptual tasks like speech, language, and vision, as well as complex internal representation tasks such as text-based information retrieval and natural language processing.

Understanding these paradigms is crucial for leveraging the full potential of DL architectures in various applications. The diverse methodologies of DL underscore its transformative potential, setting the stage for a deeper exploration of the core architectures and technologies that underpin these learning paradigms.

2.1.2 Core Architectures and Technology

Through its transformative capability, DL models have demonstrated unprecedented accuracy in several tasks, through architectures such as Multi-Layer Perceptrons (MLPs), Recurrent Neural Networks (RNN) and, of course, CNN. These architectures form the backbone of many DL-specific frameworks and are fundamental to advancing the field of AI.

Each of the aforementioned architectures is designed to deal with different types of data and modeling challenges, making them collectively fundamental to advancing the field of AI. For instance, MLPs are a foundational DL architecture capable of approximating virtually any continuous function to a high degree of accuracy, making them highly versatile for a wide range of predictive modeling tasks, including classification and regression.

CNN, a cornerstone of DL architectures, are specific designed to process data in a grid-like topology, such as images. The use of CNN has allowed the accuracy and efficiency of detection tasks to be increased [19]. Often employing multiple convolutional layers, these are excellent at extracting a variety of features and creating feature maps. These maps are pooled, activated, and combined for output generation, allowing CNN to effectively recognize complex features and patterns in medical data [20]. This ability is essential for preserving spatial relationships in 2D images inputs, as well as 3D data with minor modifications [21]. This demonstrates the versatility of CNN in adapting to different medical imaging modalities and their specific visual features, offering significant benefits in clinical contexts [22].

RNNs are known for the ability to handle and process series of data. Their capability to use their internal state (memory) to process sequences of input data makes them suitable for a variety of problems, such as time-series data analysis, text recognition, and scenarios where context is crucial across sequences. These networks have shown exceptional prowess in identifying patterns and features directly from raw images. The ability to learn and extract multiple features from raw datasets allows, in many cases, a dramatic increase in performance in a wide range of applications [23].

2.1.3 Broader applications of Deep Learning

DL significantly impacts numerous domains, demonstrating its versatility and robust architectures, capable of learning from complex and high-dimensional data. Its contributions to these fields highlight its potential in a wide range of applications and uses cases.

In Natural Language Processing, DL models have revolutionized machine translation and analysis through techniques such as tokenization. Natural Language Processing models, particularly those using RNNs and, more recently, transformers, have dramatically improved the accuracy and fluency of machine translation. Sentiment analysis has also benefited from DL's ability to understand the nuances of human language, enabling more sophisticated interpretation of text data [24].

Visual Recognition is another common DL application, where the goal is image classification and object detection. In this field, CNN have become the backbone of image recognition tasks, capable of identifying objects, faces, and scenes with remarkable accuracy. In video recognition, the combination of CNN for spatial understanding and RNNs or 3D CNN for temporal dynamics has enabled advanced applications such as activity recognition and video classification [25].

Autonomous systems such as self-driving cars, robots, drones and other autonomous vehicles are also among the many applications of DL. DL is at the heart of autonomous navigation systems, where it is used for tasks such as object detection, scene segmentation, and path navigation. The ability to process and interpret visual data in real time is critical for the safety and efficiency of these systems. However, in this context, DL techniques may be inherently unpredictable, so new or alternative methods are needed to provide strong assurance guarantees. The use of compositional verification [26] or deep reinforcement learning are alternative approaches [27].

The application of DL across various fields has demonstrated its versatility and impact, using several technologies under through different paradigms. However, the broad context of DL technologies necessitates a focused examination when considering their historical influence on healthcare and medical imaging. To truly evaluate AI's impact, particularly the DL subfield, a deeper analysis of complex medical data is required. Within the healthcare community, several subareas have been positively affected by DL, especially in MIA.

2.2 Medical Image Analysis Development

2.2.1 Initial Journey

The path of medical diagnostics has been intrinsically linked to the evolution of MIA, a field at the intersection of healthcare and technology that has undergone many changes over the years. The ability to visualize the internal state of the body has revolutionized diagnosis, surgical planning, and disease monitoring, making MIA a cornerstone of modern medicine.

Although it remains a growing field, the sector was originally driven by rule-based systems. These systems were simple in their approach, with predefined rules and algorithms for tasks such as image processing (edge detection, region growing), mathematical modeling (fitting lines, circles and ellipses), but limited in capabilities compared with to explicit programming algorithms [28]. The explicit algorithms required precise definitions and could not easily accommodate the variability inherent in medical imaging, such as variations in patient anatomy or image quality.

As the field of medical imaging advanced and the challenge of deciphering medical images arose, ML techniques emerged, a step that marked the initial use of supervised learning techniques [19]. Early examples included Active Shape Models for recognizing and tracking shapes in images, and Atlas methods for building reference models from sets of training images [29],

useful for anatomical segmentation and registration tasks. Alongside these initial approaches, the field also saw the introduction of statistical classifiers, such as Support Vector Machines and Decision Trees. These techniques focused on the process of feature extraction, manually identifying and coding algorithms to detect specific characteristics or patterns in the data, with the purpose of being fed into classifier models, mainly in disease detection and diagnosis tasks [28].

The reliance on expert-crafted features meant that these systems often required extensive medical domain knowledge, and their performance varied significantly based on the data they were trained on [28]. The manual process on extracting handcrafting feature was limited by being time-consuming, manually complex (highly dependent on domain expertise, even bigger on medical areas), and highly dependent of the quality of the features extracted (generalization) [22]. The development and refinement of techniques and algorithms were significantly propelled by advancements in hardware technology, particularly Graphics Processing Units (GPUs). These GPUs enhanced the ability to process mathematical operations, paving the way for more complex and efficient approaches [29]. The advent of efficient parallel computing, catapulted not only CNN but DL architectures in general into the spotlight in both computer vision and MIA [12].

Driven by this advance of technology, the DL era came into sight, mainly with the purpose of automation of this difficult manual feature extraction process. This paradigm shift mainly represents the move from manual to automated, data driven feature learning, effectively reducing the expert intervention and with the ability to handle varied and complex/specific data on different clinical and imaging contexts. This capability has made DL indispensable for tasks ranging from segmentation, image classification, object detection, to registration and, more importantly, landmark detection in medical images [22].

2.2.2 Importance and Challenges

Improving the automation of medical diagnosis depends on several factors. In addition to the inherent bureaucracy involved in the provision of medical images by the hospital and the regulatory scenarios to guarantee patient privacy, factors such as the variability and complexity of the data present significant challenges. Medical imaging encompasses a wide range of modalities, such as MRI, CT, X-Ray, and Ultrasound (US). Each one with distinct characteristics and diagnostic utilities. The diversity of acquisition equipment and protocols further contributes to variability, as different machines and settings can produce images with varying resolutions, contrasts, and noise levels. Additionally, the differences intrinsic to each patient such as age, gender or pathological condition, make it difficult to automate the process. This variability presents a challenge for DL models, which need to learn generalisable features that are robust to these differences.

As with any ML paradigm, the scarcity of labelled data is another challenge facing researchers. In high-stakes environments such as medical diagnosis and treatment planning, the importance of carefully curated and annotated datasets cannot be overstated, yet high-quality annotated medical images, which are crucial for training supervised DL models, are often scarce [30]. The general need for high accuracy and reliability is always fundamental to correct model operation. Given the high stakes involved in medical diagnosis and treatment, DL models need to be exceptionally accurate and reliable. Dataset preparation is therefore not just a preliminary step, but a critical and laborsome stage in the development of accurate and reliable models. The consensus among researchers is that this phase is fundamental, with the understanding that the precision and reliability of DL outcomes are clearly linked to the integrity of the dataset [13]. Ensuring that datasets are comprehensive and reflect the diversity and complexity inherent in medical imaging is essential to minimize the risk of false positives and negatives, thereby increasing the overall effectiveness of automated systems.

2.2.3 Advancements in Deep Learning for Medical Image Analysis

As already mentioned, DL has brought significant advancements, and specific to various tasks in MIA, transforming the way medical data is interpreted and utilized. The way specific tasks are approached is quite comprehensive, as seen DL realm has a broad number of paradigms and technologies perfectly adjusted to fit any assignment. There are several tasks worth mentioning, such as object detection, which involves identifying and localizing specific objects such as tumors, lesions, or organs within medical images.

Segmentation is another task, involving partitioning images into distinct regions representing different tissues, organs, or pathological areas. Accurate segmentation is crucial for numerous clinical applications, such as tumor delineation in oncology. Encoded/Decoder CNN-based architectures, have become the gold standard for medical image segmentation due to their ability to capture fine details and spatial hierarchies.

Additionally, tracking involves monitoring the movement or changes of objects over time, which is vital in longitudinal studies, such as tracking tumor growth. Architectures based on RNN are employed to achieve precise tracking in medical imaging.

Classification problems, involving categorizing images or specific regions into predefined classes (e.g., benign vs. malignant lesions), benefit greatly from CNN.

Image registration is another tasks, essential to align images from different modalities or time points to a common coordinate system, essential for accurate comparison and analysis.

Finally is landmark detection, crucial to identify the location of key anatomical landmarks in medical images, serving as a reference for diagnostic procedures and surgical planning [31].

2.2.4 Pioneering DL Architectures and Techniques

Overall, CNN are foundational to most DL-based MIA tasks. Their capability to learn hierarchical features from raw pixel data makes them ideal for the majority of the tasks. Key CNN-based architectures such as VGG, Residual Networks (ResNet) and U-Net have demonstrated remarkable performance across these tasks. For example, VGGNet [32] explores the idea that network depth is crucial for good performance, utilizing small convolution filters (3x3) throughout its architecture [33]. Residual networks [34], such as ResNet, use skip connections (Figure 2.1) that allow the flow of gradients directly through the network, mitigating the vanishing gradient problem. Including this kind of skip links might benefit the gradient calculation, as regularisation will skip any layer that degrades the performance of the architecture, facilitating the learning process [35].

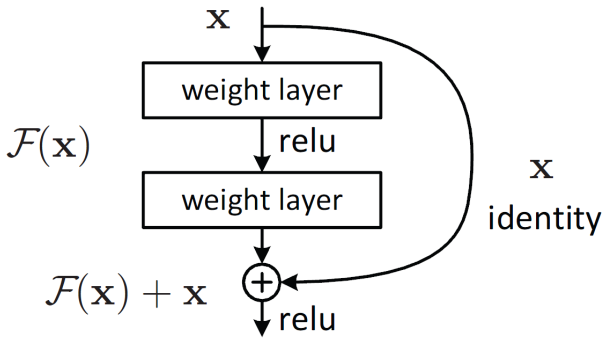


Figure 2.1: Building block from residual learning [34].

Other key techniques in DL have propelled these tasks forward significantly. Transfer learning, for instance, leverages pre-

trained models on large datasets like ImageNet and fine-tunes them for specific MIA tasks. This approach dramatically reduces the need for extensive annotated medical datasets and accelerates model development. Attention mechanisms, which enable models to focus on the most relevant parts of an image, have been integrated into various DL architectures to enhance accuracy and interpretability. Techniques like self-attention and attention gates (Figure 2.2) that integrate spatial attention information to guide the network in focusing on significant areas within a specific input, significantly contribute to performance improvements across multiple tasks.

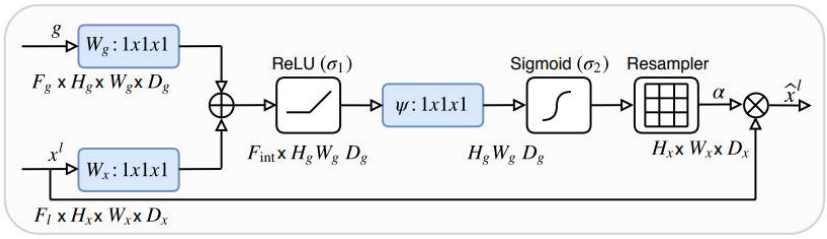


Figure 2.2: Attention gate [36].

Other common and frequently used tool is data augmentation, for dealing with data scarcity, subject to the details and specifications required to the context itself. It consists of artificially increasing the variability of training datasets by applying a series of transformations to existing data. These transformations should logically reflect variations that might occur in clinical settings to ensure that the augmented data is realistic and useful for training robust models. The application of data augmentation suggest an innovative approach, such as Generative Adversarial Networks (GANs). These networks (Figure 2.3) are used for tasks such as data augmentation, synthetic image generation, and anomaly detection. By generating realistic medical images, GANs help augment training datasets, improving model robustness.

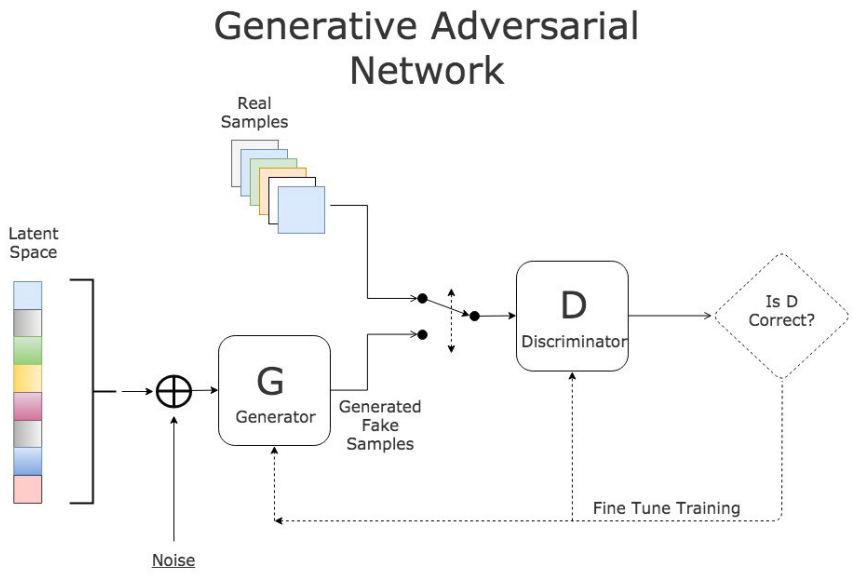


Figure 2.3: Generative Adversarial Network [37].

2.2.5 Research Path

It has been proven that DL models are at the forefront of MIA research, providing innovative solutions that not only outperform traditional methods in terms of accuracy and efficiency but also critically influence the outcomes of clinical interventions and diagnostics [31]. Given the broad scope of MIA, several interconnected tasks are addressed by DL technologies, including object detection, segmentation, tracking, classification, and registration. These tasks, while distinct, share underlying DL techniques and advancements that drive progress across the entire field. Each task leveraging these technologies in unique ways to meet its specific requirements.

Despite the considerable advancements made in these areas, it is essential to refine research into specific tasks to achieve more targeted and effective solutions. Entering the thesis content, landmark detection, is a critical task within MIA. Through the use of large datasets and advanced neural network architectures, the researchers were able to develop models capable of accurately identifying and locating key anatomical points. It required a focused research to address its unique challenges, such as variations in anatomy, pathology, and imaging modalities. Justifying that, on overall, it is impossible to determine the correct architectures or technology for this type of problems, as these are study specific and each pipeline is dependent on assertive and particular components.

Despite the significant progress made in applying DL to landmark detection in knee medical images, the journey towards fully automated, accurate and standardized diagnostics is ongoing. Future research will likely focus on refining the methodologies and techniques discussed. As researchers continue to push the boundaries of what DL can achieve in medical diagnostics, the ultimate goal remains clear: to provide clinicians with reliable tools that support early detection, accurate diagnosis, and effective treatment planning. Therefore, it is crucial to conduct a thorough review that addresses the existing literature on the landmark detection task, particularly for guiding future research in this critical area.

Chapter 3

Review on Deep Learning approaches for Landmarks Detection on Medical images

This review presents an overview of the recent DL approaches used for landmark detection in medical images. It aims to review the body area and respective amount of landmarks detected, as well as the dataset, DL architecture, and evaluation metrics used. The discussion section highlights the strengths and limitations of the reviewed studies. It determines not only the aspects that should be considered for anatomical landmarks detection on medical images but also as potential areas for future research based on DL algorithms. The conclusion section summarises the main findings of the review and highlights the importance of DL for landmarks detection in medical images. Overall, the reviewed studies demonstrate the effectiveness of DL techniques for landmarks detection in various medical imaging modalities and body areas. However, further research is needed to improve the robustness and generalizability of these methods.

3.1 Methodology

The literature search was performed according to the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA). The studies included in this systematic review were searched in the following four databases: Scopus (search field: “Article Title, Abstract, Keywords”), IEEE (search field: “All Metadata), Web Science (search field: “All Fields”) and PubMed (search field: “All fields”). The following combination was applied: ”(Deep learning OR cnn) AND (landmarks OR ”landmark detection”) AND ”medical image”.

The reference list of all the relevant studies found was checked. Only those studies that met all the eligibility criteria were included in this review. The inclusion criteria were: (1) anatomical landmark detection task performed, (2) use of DL for landmarks detection, (3) use of medical modality images, and (4) human anatomy related studies. The exclusion criteria were: (1) lack of landmark detection tools or tasks, (2) non-use of DL in architecture, (3) type of image not belonging to any type of medical modality, and (4) use of medical modality not related to the human body.

3.2 Results

The search strategy (Figure 3.1) conducted in the aforementioned databases resulted in 392 studies. After removing the duplicate papers (106 studies), 286 studies remained for screening, of which 163 were removed based on title and abstract. A total of 123 full-text studies were assessed for eligibility and, according to the exclusion criteria, the studies were removed as follows: impossible to read (foreign language or without access) (n=15), not original research (n=6), absence of landmarks

detection task (n=10), landmarks detection is not the main focus (n=17), study on metrics to improve landmarks (n=8), landmark content not enough (n=19), area of study not relevant (n=5). A total of 43 studies were included.

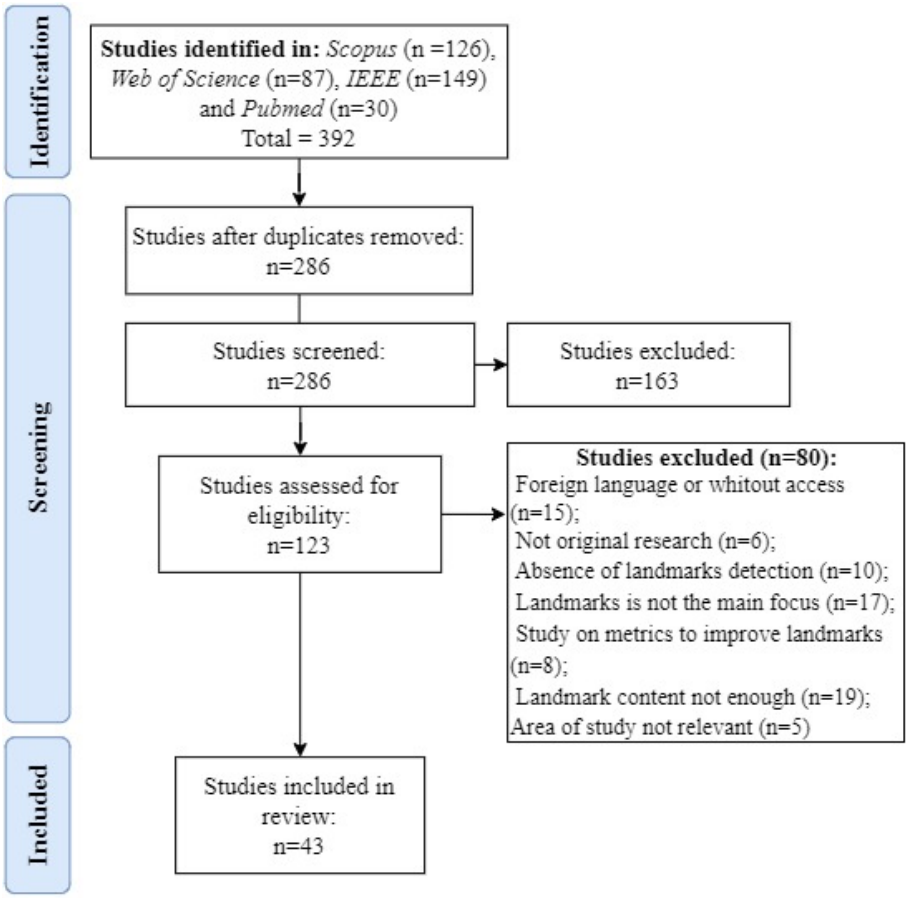


Figure 3.1: Flow diagram of search strategy based on PRISMA.

Each included study was analysed, and the following specifications were extracted: body area (area and number of landmarks), dataset (modality, size, and dimension), architecture (frameworks, components, backbones models, and variants) and evaluation metrics. The extracted data are shown in Table 3.1.

Table 3.1: Characteristics of the included studies (Body Area [**L**], **L**: Number of anatomical landmarks, Dataset [**Z**], **Z**: size, **MO**: Modality, **DIM**: Dimension), **NA**: Not Available

Study (Year)	**Body Area**	**Dataset**	**Architecture**	**Evaluation Metrics**
Yang et al. (2015) [38]	Knee [7]	Osteoarthritis Initiative [50 volumes], **MO:** MRI, **DIM:** 3D	CNN	Mean Error (ME): 4.69 ± 2.30 mm
Zheng et al. (2015) [39]	Neck [1]	Head-neck images [46 to 1181], **MO:** CT, **DIM:** 3D	Multilayer perceptron (MLP) neural network (CNN with different optimizations)	ME: 2.64 ± 4.98 mm
Le et al. (2017) [40]	Heart [6]	81 patients [234 volumes], **MO:** MRI, **DIM:** 3D	3D extension ENet architecture	Median Error: 8.8 mm, Dice: 0.83
Liefers et al. (2017) [41]	Head (Eyes-Fovea Centers) [1]	European Genetic (EUGENDA) [880 volumes], **MO:** Optical Coherence Tomography (OCT), **DIM:** 3D	Fully Convolutional Neural Network (FCNN)	Accuracy: 96.9 % (1.5 mm diameter range), ME: 73 ± 112 µm
Li et al. (2018) [42]	Obstetrics [10]	72 volumes, **MO:** Ultrasound (US), **DIM:** 3D	Patch-based Iterative Network (PIN)	Localization error (LE): 5.59 ± 3.09 mm
Goutham et al. (2019) [43]	Head (Cephalometric analysis) [7]	ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	U-Net Modified	Successful Detection Rate (SDR) [ranges 2 (65.13%), 3 (77.24%), 4 (84.69%) mm], Dice: 88 %
Liu et al. (2019) [44]	Pelvis [6]	Constructed of pelvis X-ray images [9813], **MO:** X-Ray, **DIM:** 2D	FR-DDH Network (Faster R-CNN elements with ResNet)	Point to point Error (PE): 1.244 mm, SDR [ranges 1.5 (71.41%), 2 (83.85%), 3 (95.18%) mm], MAE: 1.24 mm
Qian et al. (2019) [45]	Head (Cephalometric analysis) [19]	ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	CephaNet (Faster R-CNN modified)	Detection Accuracy [ranges 2 (82.5%), 2.5 (86.2%), 3 (89.3%), 4 (90.6%) mm]
Tiulpin et al. (2019) [46]	Knee [16]	From Hospital (Oulu University Hospital, Finland), A [81 images] and B [107 images], **MO:** X-Ray, **DIM:** 2D	Partial modern hourglass-like encoder-decoder (entry block, hourglass block, and output block)	Percentage of Correct Keypoints (PCK) [ranges 1 (14.60 ± 4.83%), 1.5 (47.52 ± 2.20%), 2 (78.88 ± 0.88%), 2.5 (93.48 ± 0.44%) mm] (A), [ranges 1 (11.24 ± 0.34%), 1.5 (44.98 ± 0.68%), 2 (75.12 ± 2.71%), 2.5 (92.11 ± 0.34%) mm] (B)

Study (Year)	Body Area	Dataset	Architecture	Evaluation Metrics
Xu et al. (2019) [47]	Obstetrics [15]	From 3T Skyra scanner [1705], **MO:** MRI, **DIM:** 3D	Proposed CNN-based heatmap regression (with 3D Hourglass network) including Markov Random Field (MRF) and 3D U-Net	PCK: $<$5 mm (77.8%) and $<$10 mm (96.4%), ME: 4.47 mm, Median Error: 3.42 mm
Eslami et al. (2020) [48]	Head (Mouth) [21]	Midsagittal MRI augmented [6138], **MO:** MRI, **DIM:** 3D	Flat-net (study-specific CNN architecture)	Root Mean Squared Error (RMSE): 0.36 cm
Liu et al. (2020) [49]	Pelvis [6]	Provincial Children's Hospital [10000 images], **MO:** X-Ray, **DIM:** 2D	PN-UNet (PointNet and U-Net based)	PE: 0.9286 mm, SDR [range 2.5 (90%), 4 ($\approx$100%) mm], Recall (92.86%), Precision (96.02%), F1 score (94.68%)
Ma et al. (2020) [50]	Neck [1]	Weill Cornell Medicine/New York Presbyterian Hospital [263 volumes], **MO:** CT Scans, **DIM:** 3D	Loc-Net (3D U-Net based)	Euclidean Distance: 2.81 $\pm$ 2.37 mm
Mozaffari et al. (2020) [51]	Head (Mouth-tongue) [5, 10, 15, 20, 25, and 30]	UOttawa [2000], **MO:** US, **DIM:** 2D	TongueNet modified light-version DNN (study-specific CNN architecture)	Mean sum of distance: 4.87 pixels
Noothout et al. (2020) [52]	Head (Cephalometric analysis) [8]	ISBI 2015 Cephalometric [400], Private Dataset (University Medical Center Utrecht) [672 scans], Private Dataset (Hospital Gelderse Vallei, Ede, The Netherlands) [61], **MO:** CT, **DIM:** 2D and 3D	ResNet34	Euclidean Distance (median): 1.15 mm, SDR [ranges 2 ($>$80%), 2.5 ($\sim$90%), 3 (92-95%), 4 (95-100%) mm]
Qian et al. (2020) [53]	Head (Cephalometric analysis) [19]	ISBI 2015 Cephalometric [400], **MO:**X-Ray, **DIM:**2D	CephaNN (2 U-Net shape subnets with ResNeXt as backbone) with multi-attention mechanism and RE loss	Mean Radial Error (MRE): 1.15 mm, SDR[ranges 2 (87.61%), 2.5 (93.16%), 3 (96.35%), 4 (98.74%) mm]
Ren et al. (2020) [54]	Head (Mouth-dental surface) [8]	Unnamed, different subjects [108], **MO:** X-Ray, **DIM:** 2D	RetinaNet (ResNet based)	Loss (mean loss: 0.0458 pixels)
Bekkouch et al. (2021) [55]	Pelvis (Hip) [16]	[140], **MO:** MRI, **DIM:** 3D	Feature Pyramid Networks (FPN) with Region Proposal Networks (RPN), 3D U-Net and Fast R-CNN	ME: 1.74 $\pm$ 1.04 mm, Percentage of landmarks predicted [ranges 1 (0%), 2 (63%), 3 (78%) mm]

Study (Year)	Body Area	Dataset	Architecture	Evaluation Metrics
Belikova et al. (2021) [56]	Head (Mouth) [1]	In AVI and DICOM [13 sequences], **MO:** US(video), **DIM:** 2D and 3D	3D U-Net	Mean squared Error (MSE): 0.0007 ± 0.0001, Dice: 0.682 ± 0.180, Euclidean Distance: 2.137 ± 2.863 mm
Bhatkalkar et al. (2021) [57]	Head (Eyes-Optic Disc and Fovea Centers) [1]	G1020 [1020], Messidor and IDRiD grand challenge fundus [413], **MO:** Color fundus Images, **DIM:** 2D	FundusPosNet (U-Net model backbone)	Euclidean Distance: 16.760 (IDRiD dataset)
Chen et al. (2021) [58]	Head (Cephalometric analysis) [18]	[80 sets], **MO:** Cone-beam Computed Tomography (CBCT), **DIM:** 3D	3D faster R-CNN with location refinement using a MS-UNet	Localization Accuracy: 0.89 ± 0.64 mm
Danilov et al. (2021) [59]	Heart [11]	Research Laboratory for Processing and Analysis of Big Data [35 series], **MO:** Transcatheter Aortic Valve Implantation (TAVI) video imaging series, **DIM:** 3D	MobileNet V2, ResNet V2, Inception V3, Inception ResNet V2 and EfficientNet B5	Precision: 0.97, Recall: 0.97, F1 score (Macro: 0.93, Micro: 0.97), Accuracy: 0.97, MAE: 0.046, MSE: 0.006, RMSE: 0.079
Kang et al. (2021) [60]	Head (Cephalometric analysis) [19] and Hand [37]	Digital Hand Atlas dataset [895] and ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	U-Net Based	PE (Mean: 0.71 ± 1.26 mm) (Digital Hand Atlas dataset), Error detection rate [ranges >2 (13.26%), >2.5 (7.89%), >3 (4.91%), >4 (1.79%) mm] (test dataset 1), [ranges >2 (25.21%), >2.5 (18.05%), >3 (12.95%), >4 (6.79%) mm] (test dataset 2)
Kwon et al. (2021) [61]	Head (Cephalometric analysis) [19]	ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	DeepLabv3 based	MRE: 1.12 mm and Successful Decision Rate [ranges 2 (86.91%), 2.5 (91.44%), 3 (94.21%), 4 (97.68%) mm], Successful Classification Rate (SCR): 85.19% (test dataset 1), 81.96% (test dataset 2)
Lang et al. (2021) [62]	Head (Mouth-dental surface) [68]	Real-patient [77 sets (dental surfaces)], **MO:** CBCT, **DIM:** 3D	DLLNet (MeshSegNet as backbone)	LE: 0.42 mm, RMSE: 0.372 ± 0.234 mm, Misdetection rate: 0% (range NA)
McCouat et al. (2021) [63]	Pelvis (Hip) [4]	From Oxford University as part of FAI Trial [375], **MO:** X-Ray, **DIM:** 2D	U-Net and Stacked Hourglass network	LE: 1.965 ± 1.598 mm

Study (Year)	Body Area	Dataset	Architecture	Evaluation Metrics
Palazzo et al. (2021) [64]	Head (Cephalometric analysis) [5]	AirwaysSet [19 scans] and anonymized scans [50 scans], **MO:** CT, **DIM:** 2D and 3D	3D variant of Tiramisu architecture (FCN), Long short-term memory (LSTM) layers and final CNN Network	LE: 0.85 mm (AirwaysSet dataset)
Reddy et al. (2021) [65]	Head (Cephalometric analysis) [19] and spine [68]	ISBI 2015 Cephalometric [400] and spinal anterior-posterior [338], **MO:** X-Ray, **DIM:** 2D	Local-appearance network and global-context network (FCN)	MRE: 1.26 mm (test datasets 1 & 2), SDR[ranges 2 (81.85%), 2.5 (87.73%), 3 (92.06%), 4 (96.51%) mm] (test datasets 1 & 2)
Tabata et al. (2021) [66]	Head (Cephalometric analysis) [19]	ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	Faster R-CNN concept of CephaNet, using ResNet50 with FPN	ME: 0. 901 mm, Success rate [range between 0, 2.5]: 97.53%
Torres et al. (2021) [67]	Head (Cephalometric analysis) [9]	constructed set of MRI of infant's heads [1250 models], **MO:** MRI, **DIM:** 2D and 3D	CNN with VGG19	ME: 4.5 ± 4.1 mm, Precision Rate: [0 ($\sim$85%), $>$10 (decline starts) mm]
Wang et al. (2021) [68]	Heart [7]	Hospital of USTC [1019 echo cine series], **MO:** US(Echo),**DIM:** 3D	ResNet enconder (RNN as a frame identification), specific CNN decoder	average frame difference: 1.56 ± 1.35 frames, PE:5.65 ± 7.60 mm, Identification Rate: 0.833
Zhang et al. (2021a) [69]	Head (Skull) [37]	West China Hospital of Stomatology [1,005], **MO:** X-Ray, **DIM:** 2D	Key Pairwise Relational Reasoning Network (KPRR)	MRE: 1.05 mm, SDR[ranges 2 (89.31%), 2.5 (93.86%), 3 (96.47%), 4 (98.59%) mm]
Zhang et al. (2021b) [70]	Head (Mouth) [4]	[59810 images], **MO:** X-Ray (videofluoroscopy), **DIM:** 2D	Two-stage networks (CNN with ResNet blocks)	Mean Localization Distance: 4.20 ± 5.54 pixels, Intraclass correlation coefficient (ICC): 0.9
Ahmed et al. (2022) [71]	Head (Skull-Brain) [1]	MRI GARD [326], MRI ADNI [351], **MO**: MRI, **DIM**: 3D	Hough regression network (HRN), with coarse prediction network, Fine tuning network (FTN) with Siamese verification network (SVN)	LE: 1.55 ± 0.61 mm (ADNI dataset)
Caspersen et al. (2022) [72]	Heart [6]	Subset of SCAPIS [500], **MO:** computer tomography (CT) Scans, **DIM:** 2D and 3D	VGG16 and ResNet50	RMSE: 10.34 ± 4.01 mm, Accuracy: 97.1%, Precision: 90.6%, Recall: 89.7%, MAE (phase 1): 2.22 ± 5.41 slices
Du et al. (2022) [73]	Head (Cephalometric analysis) [19]	Peking University, Public Dataset [4396] and Private Dataset [4000], **MO:** X-Ray, **DIM:** 2D	HRNet-18; compared backbones: DeiT, EffientNet, SEResNeXt and ResNeXt	MRE: 1.01 ± 0.85 mm, SDR[ranges 2 (80.16%), 2.5 (86.26%), 3 (90.16%), 4 (95.00%) mm]

Study (Year)	Body Area	Dataset	Architecture	Evaluation Metrics
Fard et al. (2022) [74]	Spine (Cervical) [1]	[24,419], **MO:** X-Ray, **DIM:** 2D	PoseNet	Normalized mean error: 4.75% and Failure rate: 2.77% (normal pose dataset), 3.33% (extension subset), 9.82% (flexion subset)
Jafari et al. (2022) [75]	Heart [2]	Echo video [4,493], **MO:** US (echo videos), **DIM:** 3D	U-LanD framework (Bayesian U-Net)	MAE: 1.08 ± 0.89 mm, Maximum Absolute Error: 4.66 mm, R^2 score: 66%
King et al. (2022) [76]	Head (Cephalometric analysis) [19]	ISBI 2015 Cephalometric [400], **MO:** X-Ray, **DIM:** 2D	CephaX	MRE: 1.17 ± 0.93 mm, SDR[ranges 2 (86.14%), 2.5 (91.72%), 3 (94.91%), 4 (97.96%) mm]
Leitner et al. (2022) [77]	Muscles and tendons in limbs [1]	5 US systems [66864 images], **MO:** US, **DIM:** 2D	U-Net based	RMSE: 4.89 mm, Standard Error of the Mean: 0.10 mm, MAE: NA, ICC= 0.88
SchurerWaldheim et al. (2022) [78]	Head (Eyes-Retina) [1]	B-scans [5586 volumes], **MO:** OCT, **DIM:** 2D and 3D	Prior regularization U-Net (PRE U-Net)	Euclidean distance: 0.169 ± 0.159 mm
Shankar et al. (2022) [79]	Obstetrics [6]	[1192] (TV plane [596] and TC plane [596]), **MO:** US, **DIM:** 2D	U-Net, Stacked Hourglass and HRNet	MAE: 1.98 ± 0.89 mm
Tan et al. (2022) [80]	Head (Skull-Cerebrovascular analysis) [19]	UNC public [109], UNC private [40], **MO:** Cerebral Magnetic Resonance Angiography (MRA), **DIM:** 3D	U-Net with ResNet block (Backbone of multi-task framework)	MRE: 1.81 mm, Dice: 54.25%, Accuracy: 88.95%

3.2.1 Publications per year

Figure 3.2 illustrates the number of studies published per year. According to the search strategy, only studies from 2015 onwards were found. Overall the head obtained a total of 25 studies, the heart with 5, Pelvis with 4 studies, obstetrics with 3, neck, knee and spine with 2 studies, finally hand, Muscles and tendons in limbs with 1 study.

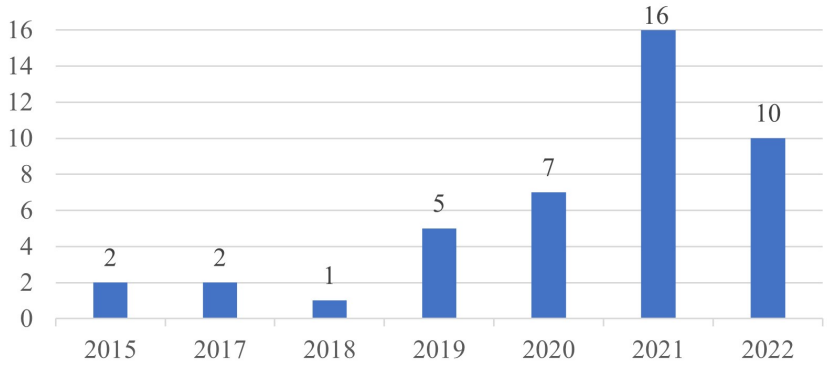


Figure 3.2: Number of studies published per year.

3.2.2 Body Area

In the reviewed studies (Table 3.1), 9 body areas can be identified (head, heart, pelvis, obstetrics, neck, knee, spine, muscles, and hand) (Figure 3.3) along with the respective diagnostic parameter to be automated.

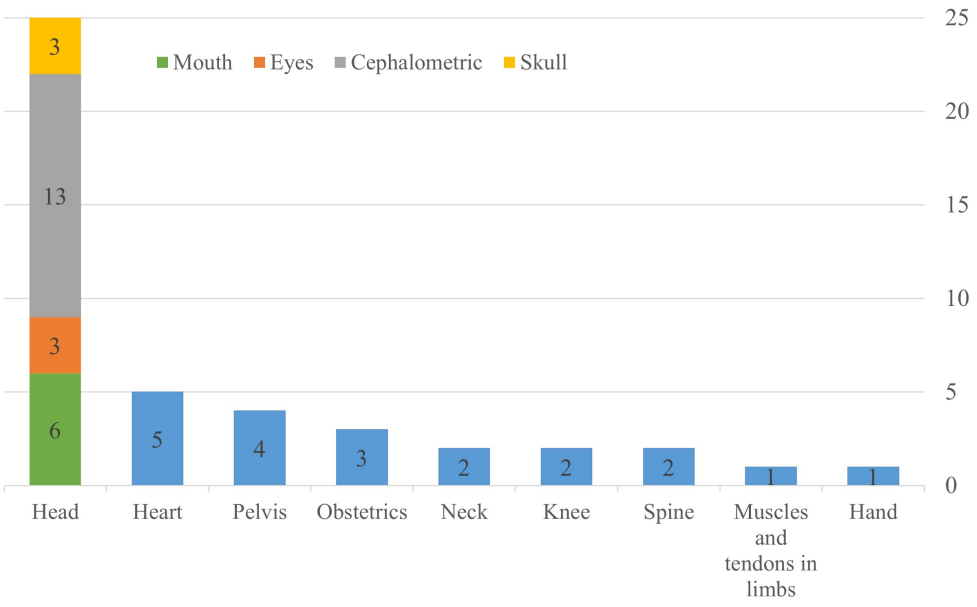


Figure 3.3: Body area distribution across studies.

Of the included studies, 25 focused on the head falling into 4 sub-areas: cephalometry, mouth, eyes, and skull. Cephalometric analysis was the main focus, with a total of 13 studies assessed. Several of these studies addressed orthodontics for oral maxillofacial surgery (7 landmarks [43] and 19 landmarks [45, 53, 66, 73, 76]) and for orthognathic surgery (19 landmarks) [61], craniomaxillofacial deformities (18 landmarks)[58], congenital birth deformities (5 landmarks) [64], the diagnosis

of cranial deformities specifically head growth evolution analysis (9 landmarks) [67], the study of bone age to diagnose growth abnormalities (19 landmarks) [60], for medical imaging analysis (8 landmarks) [52] and to surgical planning applications (19 landmarks) [65].

The mouth was examined in 6 studies, namely the dental surface and tongue, to assess orthodontics and orthognatic surgery (68 landmarks) [62] including temporomandibular joint disorders (1 landmark) [56] and osteoporosis analysis (8 landmarks) [54], tongue movement and speech (articulation) (5, 10, 15, 20, 25, and 30 landmarks) [51], swallowing disorders (4 landmarks) [70], and speech disorders and therapy related planning (21 landmarks) [48].

Eyes were the focus of 3 studies that specifically looked at the optic disc and/or foveal centres in the retina to monitor retinal disease [57], macular disease and visual impairment [78], central retinal thickness and drusen count [41] (1 landmark detected in each study).

The skull was evaluated in the last 3 studies of the head to assess anatomical abnormalities of the skull (37 landmarks) [69], cerebrovascular disease such as aneurysms and stenosis (19 landmarks) [80], and Alzheimer's and related dementias (1 landmark) [71].

Following the analysis, the heart was addressed by 5 studies focusing on aortic stenosis (11 landmarks) [59], cardiovascular diseases (7 landmarks [68] and 6 landmarks [72]), and cardiac pathologies (6 landmarks) [40]. Assessing other cardiac conditions, evaluation of ventricular function and heart valve diseases were also a parameter of focus (2 landmarks) [75].

The pelvis was the focus of 4 studies. From these, 2 aimed to assess hip dysplasia (6 landmarks) [44, 49], and the other half focused on bone pathology and pelvis morphometry (16 landmarks) [55], and femoroacetabular impingement (4 landmarks) [63].

In the medical specialty of obstetrics 3 studies were found, aimed at both fetal screening (10 landmarks) [42], and neurodevelopment, monitoring anomalies of the central nervous system (6 landmarks) [79], and estimating fetal pose (15 landmarks) [47].

The knee was analysed by 2 studies to assess the stages of osteoarthritis (16 landmarks) [46], and knee surgery planning and biomechanical analysis (7 landmarks) [38].

Spine was the subject of 2 studies. One study aided in surgical planning (68 landmarks) [65], and the other in the detection of spinal deformities (1 landmark) [74].

Neck was also the subject of 2 studies. One study addressed carotid artery bifurcation (1 landmark) [39], and the other focused on extracranial carotid artery stenosis (1 landmark) [50].

Studies on the hand, as well as on the limb muscles and tendons, with 1 study each focused on endocrine disorders (37 landmarks) [60] and on functional and behavioural aspects (1 landmark) [77], respectively.

3.2.3 Dataset

An analysis of the selected studies at Table 3.1 reveals insightful information and usability for both imaging modalities and dimensions. Furthermore, Figure 3.4 illustrates the distribution of studies by medical imaging modality, providing additional context and visual representation of the predominant techniques in the field.

It can be seen that 21 studies used 2D images as input for the DL algorithm for landmark detection. Of these, 17 studies used X-ray [43, 44, 45, 46, 49, 53, 54, 60, 61, 63, 65, 66, 69, 70, 73, 74, 76], 3 studies applied techniques on Ultrasound (US) [51, 77, 79], and 1 study used Color Fundus Images/Photography [57].

14 studies used 3D volume datasets, of which 7 used MRI [38, 40, 47, 48, 55, 71] with 1 using Cerebral Magnetic

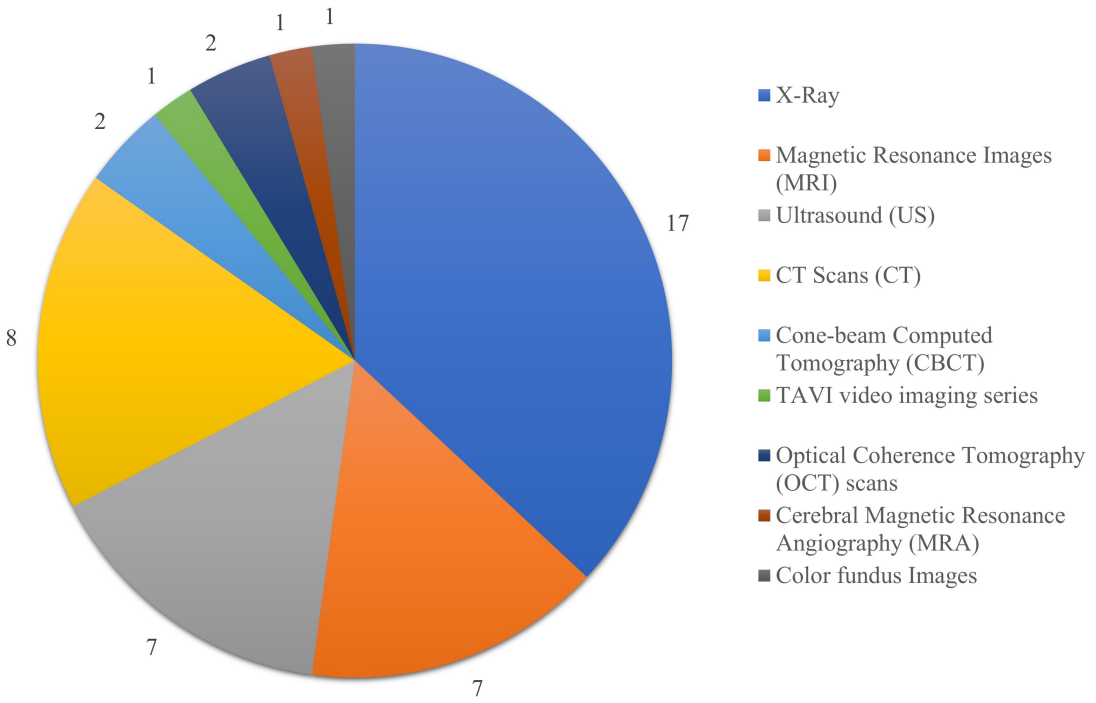


Figure 3.4: Number of studies by medical imaging modality.

Resonance Angiography (MRA) [80], 5 used computed tomography (CT) [39, 50, 58, 59, 62], of which 2 refer to Cone Beam Computer Tomography (CBCT) [58, 62] and 1 to TAVI video image series [59], 3 used US [42, 68, 75] and 1 study used Optical Coherence Tomography (OCT) scans [41].

Both 2D and 3D dimensions were used in 6 studies, 3 of which used CT scans [52, 64, 72], 1 used MRI [67], 1 used US [56], and the last used OCT scans [78].

3.2.4 Architecture(s)

A deeper analysis of the Table 3.1 reveals some common models and components in the DL domain, namely architectures associated with CNN with approaches for image classification, object detection and semantic segmentation tasks. The architecture used in each study is always dependent on the unique objective of the study itself. It is important to note the diversity of CNN architectures, each designed for specific purposes, capable of being adapted and modified to meet specific needs. These architectures/models are not only dynamic, but also offer the ability to be used individually or together, in parallel or in an integrated way, to achieve a larger, more efficient system. Finally, it is important to note that these DL models/architectures are commonly used in different fields and tasks such as computer vision, image processing, medical image segmentation, and object detection [81].

U-Net architecture was mentioned in 15 studies. This versatile model, known for its use in image segmentation tasks, consists of a contractive path that captures context, and an expansive path that enables precise localization. The architecture uses a combination of convolutional layers, maximum clustering layers, and oversampling layers to increase the spatial resolution of the feature maps [82].

Goutham et al (2019) [43] explored a DL framework inspired by the U-Net architecture for landmark detection. This adapted

U-Net model is distinguished by its lightweight design. Leveraging the architecture "downsampling-upsampling" technique, enables the network to capture features across multiple scales. Xu et al (2019) [47] used a 3D U-Net, compared it with a two-stage 3D hourglass network. Liu et al (2020) [49] introduced Pyramid Non-local U-Net, based on 2D U-Net, transforming landmark detection into a segmentation task by focusing on the local neighborhood of each landmark. Loc-Net, the approach explored by Ma et al (2020) [50], is based on a 3D U-Net, specifically a shift-equivariant network that combines a U-Net and a center-of-mass layer to produce predicted landmark coordinates. CephaNN, a multi-head attention neural network combined with attention mechanisms, with integrated U-Net shaped subnets for diverse feature learning, was used by Qian et al (2020) [53]. Bekkouch et al (2021) [55] used a U-Net to compare with its multi-stage architecture Feature Pyramid Networks (FPN) with Region Proposal Networks (RPN). Belikova et al (2021) [56] combined a 3D U-Net, responsible for spatial feature extraction, with a Long Short-Term Memory (LSTM) module, effectively tracking temporomandibular joint movements in US videos. Bhatkalkar et al (2021) [57] used a FundusPosNet, a DL model based on an encoder-decoder U-Net architecture. Chen et al (2021) [58] demonstrated a two-stage phase architecture where the first part used an extension of a 3D Faster R-CNN and the second part included a Multi Scale U-Net (MS-UNet), a lightweight variant of the 3D U-Net, which refined landmark localization through heatmap regression. Kang et al (2021) [60] employed a U-Net based architecture with an attention module that helps to filter features propagated by the skip connections, thereby improving the focus on relevant areas within the images. In addition, an anatomical context loss function was used, which takes into account the statistical distributions of distance and angle between landmarks to calculate the loss. McCouat et al (2021) [63] experimented two approaches, a simple 2D U-Net and a stacked Hourglass Network. Reddy et al (2021) [65] presented a DL architecture combining a local-appearance network, enhanced with U-Net and Convolutional Block Attention Modules, with a global context network using dilated convolutional layers trained to predict landmark locations as heatmaps. PoseNet by Fard et al (2022) [74] is an encoder-decoder based network (inspired by 3D U-Net), built on heatmap-based regression models over coordinate-based methods for landmark detection. Bayesian U-Net (BU-Net), a modified version of the U-Net architecture, is a Bayesian 2D image landmark detector that predicts both the locations of landmarks (via heatmaps) and the predictive uncertainty maps in the Jafari et al (2022) [75] study's U-LanD framework. Pre U-Net, developed by SchurerWaldheim et al (2022) [78], is characterized by being a prior regularization U-Net. It consists in a pixel-wise regression approach for automated fovea centralis detection enhanced by a spatial location prior that provides 3D context for more accurate predictions. Shankar et al (2022) [79] employed a classical U-Net model combined with a proposed Biometric Constraint-Based Supervision (BCS) method for precise automated placement of caliper points. Finally, Tan et al (2022) [80] implemented a modified U-Net architecture enhanced with ResNet based blocks, comprehended a multi-task approach that included heatmap regression, with auxiliary tasks for improved accuracy and handling of anatomical variations.

ResNet model was used in 12 studies. It is a type of CNN that introduces the concept of residual connections, which allows the network to learn residual functions with reference to the layer inputs, rather than trying to approximate the desired underlying mapping directly. The use of residual blocks makes it possible to train much deeper networks without suffering from the problem of vanishing gradients, a common issue in deep networks [34].

Liu et al (2019) [44] presented the FR-DDH network leveraging ResNet101 for robust feature extraction. The study employed a global-to-local DL strategy, using a RPN for pinpointing local neighborhood patches, applying Region of Interest (ROI) pooling to integrate region proposals with the feature map. On PN-UNet approach, Liu et al (2020) [49] compared its landmark detection results with a regression based ResNet101, alongside others to predict landmark coordinates. Noothout et al (2020) [52]

implemented ResNet-based FCNN, integrating simultaneous regression and classification tasks to accurately localize multiple anatomical landmarks. The aforementioned CephaNN by Qian et al (2020) [53] used ResNet (comparing ResNet50 and ResNeXt50) as a backbone in its subnets, taking advantage of its design of bottleneck block for feature extraction. This study explored another type of CNN as backbone, mentioned later in the review. Bekkouch et al (2021) [55] used a 3D ResNet backbone pre-trained on comprehensive video datasets. Danilov et al (2021) [59] compared several backbone architectures in its multi-task learning-based model. This group tested the ResNetV2 (based on ResNet), which successfully combined high prediction accuracy with real-time processing. Wang et al (2021) [68] used a ResNet50 encoder as the spatial features extractor of a Multi-task Learning Approach (with RNN for frame identification). Tabata et al (2021) [66] also mentioned the use of ResNet50, here applied as the backbone of the Faster R-CNN architecture, to improve feature extraction for accurate cephalometric landmark detection. Zhang et al. (2021b) [70] integrated both ResNet-50 and ResNet-34 structures in a two-level framework for coarse representation, followed by fine learning from sub-regions for precise vertebrae landmark localization. Caspersen et al. (2022) [72] employed ResNet50 in a cascaded DL framework, first to select suitable 2D slices from 3D CT scans, and then for accurate landmark detection in those slices. As mentioned when discussing the ResNet architecture, Tan et al. (2022) [80] mentioned an encoder enhanced with ResNet blocks. Finnaly, a novel and effective approach is considered in Zhang et al. (2021a) [69], which used a combination of ResNet (as backbone encoder, decoder, as an average up-sampling layer based on bilinear interpolation) for feature extraction and a relational reasoning network for landmark analysis.

Faster R-CNN architecture was identified in 5 studies. Known for its two-stage architecture, it first generates region proposals and then uses a CNN to classify and refine the region proposals. This architecture can be fine-tuned to learn specific features in anatomical landmark detection [83].

Liu et al. (2019) [44]'s FR-DDH network, mentioned earlier in the ResNet section, adopted key elements of the Faster R-CNN architecture, utilizing its RPN and ROI pooling mechanisms. Qian et al. (2019) [45] has improved CephaNet, a Faster R-CNN, by employing a multi-task loss function and a multi-scale training strategy, which allows the network to learn from images resized at different scales, effectively increasing the diversity of the training dataset. CephaNet addressed the challenge of abnormal landmarks —those that are misdetected or missed — through a two-stage repair strategy. This approach used a 'max-confidence' criterion and the Laplacian transformation to refine landmark predictions, ensuring more reliable and accurate detection results. Bekkouch et al. (2021) [55], in their multi-stage FPN/RPN architecture, combined the robust object detection capabilities of the Faster R-CNN, enhanced by the integration of multi-scale features from the FPN and the efficient region proposal generation from the RPN. In Chen et al. (2021) [58], their first stage model, a 3D adaptation of Faster R-CNN framework, used a 3D RPN to generate region proposals and a 3D Fast R-CNN to classify them and regress their bounding boxes. Finally, also mentioned above, Tabata et al. (2021) [66] employed a Faster R-CNN architecture, including the classical RPN to generate landmark region proposals and the object detection network to classify them and refine their bounding boxes.

Hourglass Networks, used in 4 studies, are usually specialized for pose estimation tasks, but their use in anatomical landmark detection is proving useful. This type of stacked hourglass architecture consists of multiple convolutional and pooling layers. The layers are arranged symmetrically, with an equal number in the contracting and expansive paths, and are connected through skip connections. This symmetrical structure is particularly effective for tasks requiring precise localization, such as detecting anatomical landmarks.

Tiulpin et al. (2019) [46] has taken advantage of an Hourglass Network architecture with a soft-argmax layer, complemented by advanced data augmentation and transfer learning techniques, to achieve sub-pixel accuracy in landmark detection tasks. Xu

et al. (2019) [47] and McCouat et al. (2021) [63] also explored the usefulness of Hourglass Networks, the former combining it with a Markov Random Field to refine keypoint localization, and the latter using a stacked version for heatmap analysis. Shankar et al. (2022) [79] also integrated Hourglass Networks, focusing specifically on caliper placement accuracy for key biometric measurements.

VGG, a classic CNN architecture, follows with 4 studies. It is based on an analysis of how to increase the depth of these networks. It uses small 3 x 3 convolutional filters and very deep architectures (with 16 to 19 layers, VGG16 to VGG19, respectively), which allows it to learn fine-grained features from the input image. Furthermore, it is characterized by its simplicity, since the only other components are pooling layers and a fully connected layer [32].

CephaNN model by Qian et al. (2020) [53], mentioned above, employed VGG19 as its backbone. Torres et al. (2021) [67] initialized and fine-tuned its initial layers of the multi-branch CNN architecture on the VGG19 model, employing transfer learning on initial feature extraction. Caspersen et al. (2022) [72] implemented a pre-trained VGG16, fine-tuning it to the specific dataset of his study. Liu et al. (2020) [49] used VGG19 as a method of comparison with its framework to predict landmark coordinate.

In the context of hierarchical models, where specific models complement each other, 4 studies were identified. These studies took advantage of versatile models and components that allowed the integration of more complex architectures, better suited to the specific task they were trying to solve.

Ahmed et al. (2022) [71], for example, used an architecture consisting of a sequential integration of three networks. First, the Coarse Prediction Network provides a global estimation of landmark locations. This is followed by the Siamese Verification Network, which enhances accuracy by validating the detected features. Finally, the Hough Regression Network (HRN) refines these features by focusing on local details, culminating in accurate detection of landmarks in the hippocampus. Bekkouch et al. (2021)[55] used FPN to build high-level semantic feature maps, with Region Proposal Networks RPN serving as a sliding window object detector that classifies and refines features for accurate landmark detection. Li et al. (2018) [42] used a CNN-based, iterative, multi-task learning approach, the Patch-based Iterative Network (PIN). The network consists of five convolutional layers followed by max-pooling layers, with each task — regression and classification — having three separate fully connected layers for learning task-specific features. PIN operates iteratively, refining the predicted positions of landmark through successive applications of the CNN until convergence is reached, making it particularly effective at locating multiple landmarks while taking into account their spatial relationships. Palazzo et al. (2021) [64] used a 3D Tiramisu architecture that incorporates dense blocks and residual squeeze-and-excitation layers for better feature extraction. This setup, by applying channel-wise attention, improved specificity in identifying relevant image features. Transition-down layers streamline data dimensionality, ensuring efficient analysis. A notable addition was the LSTM component, designed to encode spatio-temporal information and crucial for capturing the dynamic aspects of medical images.

HRNet, presented in 2 studies, stands out for its architectural approach. Known for its multi-scale, multi-branch structure, HRNet excels at learning detailed features at various resolutions, facilitated by interconnected branches via a feature fusion module [84].

This capability was exploited in Du et al. (2022) [73], which used HRNet-18, a variant of the model, for effective feature map extraction. Shankar et al. (2022) [79] also used HRNet, demonstrating its versatility in conjunction with the Hourglass Network. In this study, HRNet played a crucial role in analyzing the relationship between landmarks and the effectiveness of the bone condition score, showcasing the complementary use of these two architectures in the same research framework.

The following models were presented by only 1 study, out of a total of 13 studies where this was the case.

Le et al. (2017) [40] mentioned the use of a 3D ENet, which is a lightweight DL architecture that uses a combination of depth-wise separable convolutions and a lightweight structure to achieve high performance while using fewer computational resources. It also uses downsampling-upsampling techniques that allow the network to learn features at multiple scales, useful for tasks such as image segmentation [85].

Danilov et al. (2021) [59] experimented several DL approaches: MobileNet V2, the already mentioned ResNetV2, Inception V3, Inception ResNet V2, and EfficientNet B5. The experiment focused on the integration of different DL models into an multi-task learning framework and contributed to the real-time prediction of aortic valve and delivery system location. Of the architectures mentioned and not explained, MobileNet is a lightweight architecture designed to be efficient in terms of computational resources, which uses depth-wise separable convolutions and allows it to run on mobile devices and embedded systems with limited computational resources [86]. Inception is a DL architecture that uses multiple convolutional filters of different sizes to learn features at multiple scales, allowing it to deal images of different sizes and aspect ratios [87]. The EfficientNet architecture is a DL architecture designed to be efficient in terms of computational resources while maintaining high accuracy. It uses a combination of depth-wise separable convolutions and a lightweight structure to achieve high performance using fewer computational resources than other architectures. This architecture differs from MobileNet because it achieves higher accuracy by scaling the depth, width, and resolution of the network and by using squeeze-and-excitation layers [88], whereas MobileNet uses depth-wise separable convolutions to reduce computational resources while preserving accuracy blocks to enhance feature representation.

Kwon et al. (2021) [61] based the main detection framework on the DeepLabv3 architecture [89] using probability density functions. Detection was tuned with independent CNNs for greater accuracy.

Lang et al. (2021) [62] built a DLLNet, a two-stage coarse-to-fine strategy for landmark localization. Initially, tooth segmentation was performed on a down-sampled mesh model, followed by refined localization on mesh patches near the coarse results. This architecture was compared with other state-of-the-art models, such as MeshSegNet, a DL architecture based on a U-Net and designed for 3D medical image segmentation tasks [90]. Other architectures used for benchmarking were PointNet++ and PointConv. Both represent significant advances in the field of geometric DL, offering robust solutions for analyzing and interpreting 3D point cloud data [91].

King et al. (2022) [76] used the CephaX framework, based on Darknet53, the backbone of You only look once v3 (YOLOv3). This architecture is enhanced with a modified multitask loss and an attention mechanism for accurate cephalometric landmark detection (eliminating bounding box constraints). As a backbone for the YOLOv3 network, the DarkNet53 architecture is composed of several convolutional layers, batch normalization layers, and pooling layers that are stacked together to form a deep network. The architecture uses a combination of convolutional layers with small kernel sizes and large strides to reduce the spatial resolution of the feature maps, and convolutional layers with large kernel sizes and small strides to increase the number of feature maps [92].

The remaining 6 studies built their own architecture/network based on the basic properties of CNN to solve their problem, making it impossible to combine it with any other common architecture found in the other studies.

Yang et al. (2015) [38] built a CNN framework that first approaches landmark detection as a binary classification task. In this framework, each landmark is associated with 2D images, which are then labelled as 'positive' if they contain the landmark, or 'negative' otherwise. This methodology allows for an analysis focused on the presence or absence of specific landmarks in the given image slices, serving as a foundational step towards accurate landmark detection.

In Zheng et al. (2015) [39], a two-stage DL strategy is employed. The first stage, a shallow MLP was used for voxel testing across the volume, using a sliding window approach to generate candidate landmarks. This is followed by a more comprehensive classification using a deeper network with three dispersed hidden layers for refined landmark detection. The approach is further enhanced by using multi-resolution image patches to increase the robustness of detection, combining deep-learned features with Haar wavelet features via a Probabilistic Boosting Tree to significantly improve landmark detection accuracy.

Liefers et al. (2017) [41] used a fully CNN with five layers of dilated convolutions. The use of dilated filters in the network provides asymmetric spacing in the axial and lateral directions, which is beneficial for the specific task of landmark detection.

Eslami et al. (2020) [48] presented Flat-net, a specific CNN based on heatmap generation, characterized by the absence of pooling, down-sampling, up-sampling or fully connected layers. Instead, it uses convolutional layers with different kernel sizes and dilation rates.

Ren et al. (2020) [54] used and compared two approaches: a CNN based on RetinaNet [93] with two different conversion methods, global detection and local detection, to generate landmark coordinates, and a Statistical Shape Model not involved in the context of the review research. While global detection computes the coordinates of the ROI or label as the center of the bounding box with the highest confidence score (applied across the entire image), local detection focuses on a certain range estimated by the distribution of landmarks in the training set and selects bonding boxes within this range, ignoring outliers far from the ground truth.

TongueNet is a Mozaffari et al. (2020) [51] approach, described as a light version of a deep CNN specifically designed for localize landmarks on the surface of the tongue.

Reinforcement Learning is also a technology mentioned in 1 study Bekkouch et al. (2021) [55], where the authors implemented 3 reinforcement learning models. The first was the Deep Q-Network extension, which combines Q-learning with deep neural networks to deal with high-dimensional state spaces, making it possible to learn policies directly from raw pixel inputs [94]. The second model was the Deep Deterministic Policy Gradient (DDPG), an architecture that extends the previous Deep Q-Network to work in continuous action spaces. Unlike Deep Q-Network, which directly learns a value function to infer the policy, DDPG learns a deterministic policy that maps states to actions. It uses two neural networks: an actor that proposes an action given the current state, and a critic that evaluates the proposed action by estimating the Q-value (the expected future rewards). Finally , the third model was a Twin Delayed DDPG (TD3), an improvement on DDPG designed to address its overestimation bias in Q-value estimation, leading to smoother policy updates and more stable and reliable learning [95]. The reward function of the architectures in this study for the agents is based on changing the Euclidean distance between the proposed and actual landmark positions, encouraging agents to minimize this distance.

3.2.5 Deep Learning Pipelines

Analysing the architectures in Table 3.1, it was possible to distinguish various types of pipelines used by the studies, Table 3.2, highly dependent on the data used. These methodologies are a way of being able to compare the different methods that the various studies use in order to detect landmarks with better performance. The different DL technologies, being architectures and/or techniques might work on the same type of pipeline, helping comparing different data and clinical context studies. It is important to note that some studies utilized a single pipeline, while others employed a combination of pipelines, highlighting the versatility of DL technology.

Table 3.2: Deep Learning pipelines regarding study's main framework approach

Study (Year)	**Main Framework approach**	**Pipelines**
Yang et al. (2015) [38]	CNN	Classification-Based Candidate Generation , Global to Local Feature Analysis
Zheng et al. (2015) [39]	Two stage DL (1°MLP with 2° CNN)	Classification-Based Candidate Generation, Probabilistic and Regression Approaches
Le et al. (2017) [40]	3D Enet	Direct Heatmap Probability Mapping
Liefers et al. (2017) [41]	CNN	Direct Heatmap Probability Mapping
Li et al. (2018) [42]	PIN	Iterative Patch-based Refinement, Probabilistic and Regression Approaches
Goutham et al. (2019) [43]	U-Net	Direct Heatmap Probability Mapping
Liu et al. (2019) [44]	FR-DDH (ResNet)	Classification-Based Candidate Generation, Combination of Spatial Information and Image Features
Qian et al. (2019) [45]	CephaNet (Faster R-CNN)	Classification-Based Candidate Generation, Probabilistic and Regression Approaches
Tiulpin et al. (2019) [46]	Hourglass network architecture	Direct Heatmap Probability Mapping
Xu et al. (2019) [47]	Hourglass network architecture	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
Eslami et al. (2020) [48]	Flat-Net	Direct Heatmap Probability Mapping
Liu et al. (2020) [49]	PN-Unet	Combination of Spatial Information and Image Features, Direct Heatmap Probability Mapping
Ma et al. (2020) [50]	Loc-Net	Probabilistic and Regression Approaches, Direct Heatmap Probability Mapping
Mozaffari et al. (2020) [51]	TongueNet	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
Noothout et al. (2020) [52]	FCNN (ResNet)	Global to Local Feature Analysis, Probabilistic and Regression Approaches
Qian et al. (2020) [53]	CephaNN (2 U-Net shape subnets with ResNeXt as backbone)	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
Ren et al. (2020) [54]	RetinaNet	Global to Local Feature Analysis, Probabilistic and Regression Approaches
Bekkouch et al. (2021) [55]	Feature Pyramid Networks (FPN) with Region Proposal Networks	Global to Local Feature Analysis, Probabilistic and Regression Approaches (Reinforcement Learning)

Study (Year)	Main Framework Approach	Pipelines
Belikova et al. (2021) [56]	U-Net	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
Bhatkalkar et al. (2021) [57]	FundusPosNet (U-Net)	Direct Heatmap Probability Mapping
Chen et al. (2021) [58]	Faster R-CNN with MS-Unet	Global to Local Feature Analysis, Direct Heatmap Probability Mapping
Danilov et al. (2021) [59]	Multi-task learning (MTL) framework	Classification-Based Candidate Generation , Probabilistic and Regression Approaches
Kang et al. (2021) [60]	U-Net	Direct Heatmap Probability Mapping
Kwon et al. (2021) [61]	DeepLabv3	Global to Local Feature Analysis
Lang et al. (2021) [62]	DLLNet	Global to Local Feature Analysis
McCouat et al. (2021) [63]	Framework with U-Net and Hourglass Network	Direct Heatmap Probability Mapping
Palazzo et al. (2021) [64]	Multi-stage architecture (3D Tiramisu)	Global to Local Feature Analysis, Iterative Patch-based Refinement
Reddy et al. (2021) [65]	Local appearance network	Direct Heatmap Probability Mapping, Combination of Spatial Information and Image Features
Tabata et al. (2021) [66]	Faster R-CNN(ResNet50 with FPN)	Classification-Based Candidate Generation
Torres et al. (2021) [67]	Specific study (VGG)	Global to Local Feature Analysis, Combination of Spatial Information and Image Features
Wang et al. (2021) [68]	MTL framework - Enconder-decoder (ResNet-CNN) (RNN for frame identification)	Direct Heatmap Probability Mapping
Zhang et al. (2021a) [69]	Key Pairwise Relational Reasoning Network (backbone encoder as ResNet)	Direct Heatmap Probability Mapping, Combination of Spatial Information and Image Features
Zhang et al. (2021b) [70]	Two-Stage DL Framework	Global to Local Feature Analysis, Direct Heatmap Probability Mapping
Ahmed et al. (2022) [71]	Cascaded Hough Regression Networks (HRNs) and Siamese Network	Global to Local Feature Analysis, Probabilistic and Regression Approaches
Caspersen et al. (2022) [72]	Framework with VGG16 and ResNet50	Classification-Based Candidate Generation, Global to Local Feature Analysis
Du et al. (2022) [73]	HRNet with Pre-trained Model	Combination of Spatial Information and Image Features, Direct Heatmap Probability Mapping
Fard et al. (2022) [74]	PoseNet (U-Net)	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches

Study (Year)	Main Framework Approach	Pipelines
Jafari et al. (2022) [75]	U-LanD framework (BU-Net)	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
King et al. (2022) [76]	CephaX based on Darknet53	Classification-Based Candidate Generation, Iterative Patch-based Refinement
Leitner et al. (2022) [77]	U-Net-based model	Direct Heatmap Probability Mapping, Combination of Spatial Information and Image Features
SchurerWaldheim et al. (2022) [78]	Prior regularization U-Net (PRE U-Net)	Combination of Spatial Information and Image Features, Probabilistic and Regression Approaches
Shankar et al. (2022) [79]	Framework based on U-Net	Direct Heatmap Probability Mapping, Probabilistic and Regression Approaches
Tan et al. (2022) [80]	Multi-task DL network (U-Net + ResNet based)	Direct Heatmap Probability Mapping, Classification-Based Candidate Generation, Combination of Spatial Information and Image Features

The Direct Heatmap Probability Mapping is one of the pipelines, used in 25 studies. It generates heatmaps representing the probability of each pixel or voxel being the location of a landmark. This is usually achieved using models such as U-Net or modified versions of it.

Probabilistic and Regression Approaches, applied in 17 studies, is another suitable pipeline, where models are trained to predict the landmarks coordinates as continuous variables. This may involve treating the task as a regression problem where the output is the coordinates themselves or related spatial information such as distance maps. Some methods can incorporate probabilistic elements, such as Bayesian networks, to assess the uncertainty in landmark predictions, which is particularly useful in medical images of varying quality or pathological cases.

Global to Local Feature Analysis was presented in 12 studies. This pipeline starts by analyzing global features of the entire image or volume to narrow down areas of interest. This may involve identifying regions where landmarks are likely to be located based on the overall anatomy captured in the image. The method then zooms into local features for accurate landmark detection. This may involve using finer-scale image features or applying more detailed models specifically trained to detect landmarks within smaller, localized image inputs.

The Classification-Based Candidate Generation pipeline follows with 9 studies. It tipically uses a neural network to perform a voxel-wise classification across the entire volume. This global analysis aims to generate a manageable number of candidate locations by classifying each voxel based on its likelihood of being close to the landmark. This produces a coarse, broad heatmap of potential landmark locations.

The Combination of Spatial Information and Image Features was also found in 9 studies. In this approach, additional spatial information, such as anatomical priors or spatial location priors is used alongside image features to improve the accuracy of landmark detection. This could involve integrating spatial maps as additional input layers for DL models or using spatial context to inform post-processing steps. This approach helps to contextualize the landmark localization process by providing additional

guidance to the model based on known anatomical structures or expected landmark positions.

Finally, Iterative Patch-based Refinement pipeline, observed in 3 studies, uses patch-based updates to improve the accuracy of landmark localization, especially in medical volumes where computational efficiency and handling of spatial relationships between multiple landmarks are crucial. This leverages both local patch information and global anatomical relationships, providing a sophisticated method for landmark detection.

3.2.6 Evaluation Metrics and Results

Analysing Table 3.1, a variety of metrics are identified, each of which serves a distinct purpose in the evaluation process. It is important to understand the objectives behind these measures, especially in relation to the specific problems addressed, the architectures deployed, and the datasets and data input used. These metrics are categorized into regression and classification types. This distinction highlights how different architectures and pipelines necessitate the use of varied evaluation metrics to accurately assess the performance of the deep learning models. Throughout the table, the metrics are presented, along with the best results for each study's approach.

Regression metrics play a key role in prediction and localisation tasks, such as landmark detection. These metrics provide quantitative measures of a model's accuracy, allowing an assessment of how close the model's predictions are to the actual values. A total of 31 studies used a regression metric.

Of these, 7 studies reported Mean Error (ME). Yang et al. (2015) [38] obtained a ME of 4.69 ± 2.30 mm for all detected landmarks. Zheng et al. (2015) [39], combining deeply learned features and Haar wavelet obtained a ME value of 2.64 ± 4.98 mm. Liefers et al. (2017) [41] used ME to assess the landmark detection, obtaining a value of 73 ± 112 µm. The proposed CNN of [47] achieved a ME of 4.47 mm, corresponding to 1.5 pixels. In Bekkouch et al. (2021) [55], their TD3 architecture combined with FPN+RPN reported a ME of 1.83 ± 0.78 mm and combined with U-Net showed a ME of 1.74 ± 1.04 mm. Tabata et al. (2021) [66] evaluated their based Faster R-CNN architecture with 2 different test datasets, achieving a ME of 0.919 mm and 0.901 mm. Torres et al. (2021) [67] evaluated their VGG19 based CNN, obtaining a ME value of 4.5 ± 4.1 mm.

Mean Radial Error (MRE) was another common metric, mentioned in 7 studies. Qian et al. (2020) [53] evaluated their CephaNN with multi-attention mechanism on two test datasets, Test1 and Test2, and obtained MRE values of 1.15 mm and 1.43 mm, respectively. Kwon et al. (2021) [61], also using the proposed framework with two different test datasets, Test 1 and Test 2, obtained MRE values of 1.12 mm and 1.41 mm, respectively. Reddy et al. (2021) [65] also used two test datasets to evaluate their local appearance network, achieving MRE values of 1.14 mm, 1.44 mm and 1.26 mm for the first, second and both datasets, respectively. In Zhang et al. (2021a) [69], their encoder-decoder architecture obtained a MRE of approximately 1.05 mm. Du et al. (2022) [73] also assessed their architecture using two subsets of the test dataset, Test 1 and Test 1, reaching MRE values of 1.01 ± 0.86 mm and 1.33 ± 0.88mm, respectively. In King et al. (2022) [76], CephaX was evaluated with two test datasets, obtaining MRE values of 1.17 ± 0.93 mm and 1.50 ± 1.00 mm. The multi-task framework in Tan et al. (2022) [80], was tested on both public and private datasets, achieving 1.81 mm and 2.30 mm values for MRE, respectively.

Mean Absolute Error (MAE) was referred in 6 studies. In Liu et al. (2019) [44], the FR-DDH network achieved a MAE of 1.24 mm. In Danilov et al. (2021) [59], the Inception ResNetV2FT model stood out with the lowest MAE, recorded at 0.046. Caspersen et al. (2022) [72] reported a MAE value of 2.22 ± 5.41 slices for the first phase of optimal slice selection with ResNet50. In Jafari et al. (2022) [75], their U-LanD framework achieved a MAE of 1.08 ± 0.89 mm. Shankar et al. (2022) [79] reported that the U-Net model with data augmentation and BCS performed the best, achieving a MAE value of 1.98 ± 0.89

mm across the test set. Leitner et al. (2022) [77] mentioned that they had used the MAE, but no value was provided.

Euclidean Distance was used in 5 studies. In Belikova et al. (2021) [56], the 3D U-Net achieved a mean euclidean distance of 2.137 ± 2.863 mm. Bhatkalkar et al. (2021) [57] evaluated FundusPosNet using three different datasets. For the IDRiD dataset, they obtained an euclidean distance of 16.760 and 40.13 for the location of the optic disc and fovea, respectively. For the Messidor dataset, the reported mean values were 10.68 and 11.62, respectively. For the G1020 dataset, only the mean euclidean distance for the location of the center of the optical disc was reported, which was 54.59. SchurerWaldheim et al (2022) [78] evaluated their proposed prior regularization U-Net architecture by measuring the distance between the position of the fovea predicted by the model and the manually annotated ground truth, achieving an average euclidean distance of 0.169 ± 0.159 mm. In Ma et al (2020) [50], their Loc-Net obtained a mean euclidean distance of 2.81 ± 2.37 mm. Noothout et al (2020) [52] obtained a median euclidean distance of 1.15 mm for the Test1 dataset and 1.60 mm for the Test2 dataset.

Root Mean Squared Error (RMSE) was mentioned in 5 studies. Eslami et al. (2020) [48] achieved a RMSE value of 0.36 cm, equivalent to 3.6 pixels, for their Flat-Net. In Danilov et al. (2021) [59], the ResNet V2 FT architecture achieved a RMSE of 0.079. Lang et al. (2021) [62] evaluated their two-stage DLLNet, obtaining a RMSE of 0.372 ± 0.234 mm. In Caspersen et al. (2022) [72], in phase 2 and 3 of the framework, RMSE was the metric used as a benchmark to evaluate and compare the backbones with the manual annotation process. For the second phase, while VGG16 backbone reached 23.35 ± 7.70 mm, ResNet 50 achieved 11.02 ± 5.09 mm. The third phase tested the difference between ResNet50 with or without augmentation, achieving best value result using augmentation 10.34 ± 4.01 mm. Finally, in Leitner et al. (2022) [77], the proposed model achieved a RMSE of 4.89 mm and a Standard Error of the Mean (a statistical dispersion metric) of 0.10 mm.

Point to Point Error (PE) was reported in 5 studies. In Liu et al. (2019) [44], experiments with the FR-DDH network were performed with local neighbourhood patches of varying sizes (N ranging from 50 to 100). The study achieved an average PE of 1.244 mm when N = 80. Liu et al. (2020) [49] obtained an average PE value of 0.9286 mm with their PN-UNet approach. In assessing their attention module based U-Net, Kang et al. (2021) [60] used the Digital Hand Atlas dataset and two subsets from ISBI2015 dataset. On the Digital Hand Atlas dataset, employing a hybrid perturbator, the proposed method yielded a mean PE of 0.71 ± 1.26 mm. On ISBI2015 dataset, using an edge detector perturbator, one of the subsets obtained a mean PE of 11.48 ± 11.08 mm, while for the other a mean PE of 15.30 ± 12.65 mm was achieved. Tabata et al. (2021) [66] obtained an overall mean PE of 0.919 mm with its Faster R-CNN. Wang et al. (2021) [68] assessed the multi-task learning framework for landmark detection with a 5-fold cross-validation, focusing on end-diastolic and end-systolic frames. The PE values for end-diastolic and end-systolic frames were 5.64 ± 8.01 mm and 5.65 ± 7.60 mm, respectively. In addition, they examined the average frame difference, reporting results of 1.59 ± 1.34 frames and 1.56 ± 1.35 frames, respectively.

Localisation Error (LE) was also referred in 5 studies. The PIN architecture in Li et al. (2018) [42] achieved an average LE of 5.59 ± 3.09 mm. In Lang et al. (2021) [62], their two-stage DLLNet reached an average LE of 0.42 mm. In McCouat et al. (2021) [63], the LE was determined for each key-point, with an average value of 1.965 ± 1.598 mm. Palazzo et al. (2021) [64] evaluated their three subnetworks on two different datasets. On the AirwaysSet dataset, the model achieved an average LE of 0.85 mm. On the other dataset provided by the study for comparison of the framework, it achieved a LE of 0.78 mm. Ahmed et al. (2022) [71] assessed their Siamese network together with cascaded HRNs on two datasets, GARD and ADNI. In the GARD dataset, they obtained an LE of 1.70 ± 0.50 mm left hippocampus and an LE of 1.66 ± 0.49 mm for the right hippocampus. For the ADNI dataset, an LE of 1.79 ± 0.83 mm and 1.55 ± 0.61 mm were achieved for the left and right hippocampi, respectively.

Median Error was also mentioned in 3 studies. Le et al. (2017) [40] achieved a median error of 8.8 mm for all predictions

over the test set. In Xu et al. (2019) [47], the proposed CNN reached a median error of 3.42 mm. In McCouat et al. (2021) [63], the median error is referred to as the medial value, and a value of 1.64 mm was obtained for each key-point.

Mean Squared Error (MSE) was mentioned in 2 studies. In Belikova et al. (2021) [56], the 3D U-Net reached a MSE value of 0.0007 ± 0.0001. In Danilov et al. (2021) [59], the ResNet V2 Fine Tuned (FT) version model stands out with the lowest MSE value at 0.006.

Minimum Error and Maximum Error were also mentioned in 2 studies each. In Yang et al. (2015) [38], a mean value for Min and Max Errors was provided with 1.07 mm and 7.86 mm, respectively. In Bekkouch et al. (2021) [55], the TD3 architecture combined with FPN+RPN obtained a Min Error of 1.42 mm and a Max Error of 5.13 mm. When the TD3 architecture was combined with U-Net, it resulted in a Min Error of 1.14 mm and a Max Error of 3.82 mm.

The remaining regression metrics found were only mentioned in 1 study each. Mean Sum of Distance is one of these metrics, with a value of 4.87 pixels for the modified TongueNet from Mozaffari et al. (2020) [51].

Mean Localization Distance (MLD) was used in Zhang et al. (2021b) [70], to evaluate the two-stage network, using 5-fold Cross-Validation on two datasets. For Patient Data dataset, the MLD was 4.07 pixels across 5 folds. In contrast, Healthy Data dataset showed a MLD with values of 4.67 pixels. On a different testing set, an independent dataset involving 70 subjects obtained a mean MLD value of 4.20 ± 5.54 pixels.

Maximum Absolute Error was reported in Jafari et al. (2022) [75], who assessed their U-LanD framework with a value of 4.66 mm. In addition, they also used the R^2 score, a statistical measure of dispersion and reliability that measures the correlation between predicted and actual lengths of the left ventricular outflow tract, achieving a score of 66%.

Loss was mentioned in Ren et al. (2020) [54] to evaluate the RetinaNet-based approach. The lower the loss, the higher the accuracy of the landmark detection. The CNN evaluation focused on comparing two different conversion methods: RetinaNet + Global Detection, with a mean loss of 0.0645 pixels; and RetinaNet + Local Detection, with a mean loss of 0.0458 pixels. RetinaNet + Local Detection achieved better results, but the Patch-based Statistical Shape Model mentioned in the study outperformed both CNN-based RetinaNet approaches.

Finally, Normalized Mean Error was used in Fard et al. (2022) [74] to evaluate the PoseNet architecture. The dataset was divided into three pose-based categories: normal, extension, and flexion. For PoseNet with L2 loss, normalized mean errors of 4.75% (for normal), 5.21% (for extension) and 7.48% (for flexion) were achieved, while with L1 loss the results were 4.69%, 5.20% and 7.25%, respectively. The Intensity-aware Categorical loss achieved normalized mean errors of 4.38% (for normal), 4.76% (for extension) and 6.50% (for flexion).

Classification metrics are also crucial for assessing landmark detection models, offering insights into precision, recall, and overall accuracy. These metrics provide a direct measure of the model's success rate in making correct predictions by assessing its ability to identify and classify landmark points accurately. A total of 26 studies used a classification metric.

Of these, 10 mentioned Sucess Detection Rate (SDR). In Goutham et al. (2019) [43], for precision ranges of 2.0 mm, 3.0 mm, and 4.0 mm, the modified U-Net obtained a SDR of 65.13%, 77.24% and 84.69%, respectively. The FR-DDH network from Liu et al. (2019) [44], for precision ranges of 1.5 mm, 2.0 mm, and 3.0 mm, achieved a SDR of 71.41%, 83.85%, and 95.18%, respectively. Liu et al. (2010) [49] employed SDR to assess their PN-UNet, noting successful detection of approximately 90% of landmarks within a precision range of 2.5 mm, and nearly 100% within a range of 4 mm. The remaining 7 studies evaluated their architecture for four precision ranges: 2 mm, 2.5 mm, 3 mm and 4 mm. The results will be presented according to this order of precision ranges. Noothout et al. (2020) [52] tested the ResNet34 with two test datasets. For Test1 dataset, the SDR

was slightly above 80%, around 90%, between 92% and 95%, and between 95% and 100%, respectively. For the Test2 dataset, the SDR was between 70% and 75%, above 80%, around 87%, and between 90% and 95%. CephaNN with multi-attention mechanism from Qian et al. (2020) [53], for the Test1 dataset obtained SDR values of 87.61%, 93.16%, 96.35% and 98.74%, while the reported SDR for the Test2 dataset was 76.32%, 82.95%, 87.95% and 94.63%. Kwon et al. (2021) [61] assessed their DeepLabv3 architecture on two subsets of the ISBI2015 dataset. The first subset achieved SDR values of 86.91%, 91.44%, 94.21%, and 97.68%. The second subset reached SDR values of 77.16%, 84.79%, 89.21%, and 94.95%. In Reddy et al. (2021) [65], for the first test dataset, the proposed architecture achieved values of 86.28%, 91.12%, 94.81%, and 97.58%, and for the second test dataset, SDR values of 75.21%, 82.65%, 87.95% and 94.89% were obtained. For the two test datasets combined, the SDR values were 81.85%, 87.73%, 92.06% and 96.51%, respectively. For their encoder-decoder, Zhang et al. (2021a) [69] achieved results of 89.31%, 93.86%, 96.47% and 98.59%. For a private dataset, Du et al. (2022) [73] obtained SDR values of 94.67%, 96.44%, 97.34% and 98.22%. For their public dataset, while they achieved SDR values of 88.74%, 93.37%, 96.14% and 98.49% for the Test 1 subset, they reached values of 80.16%, 86.26%, 90.16% and 95.00% for the Test 2 subset. The CephaX framework in King et al. (2022) [76] achieved SDR values of 86.14%, 91.72%, 94.91% and 97.96% for the first test dataset, and SDR values of 74.58%, 81.74%, 87.26% and 94.73% for the second test dataset.

Accuracy was reported in 6 studies. In Liefers et al. (2017) [41], the detection of the fovea was correctly identified in 96.9% of the cases. In Danilov et al. (2021) [59], among the architectures tested, ResNet V2 and MobileNet V2 achieved accuracy rates of 97% and 96%, respectively. ResNet V2 FT, with fine tuning, demonstrated the high overall accuracy of 97%. Caspersen et al. (2022) [72] reported a high accuracy of 97.1% for phase 1 of the architecture, the binary classification network. Tan et al. (2022) [80] reported 88.95% and 97.25% of accuracy for public and private datasets, respectively. Detection Accuracy was used to assess Qian et al. (2019) [45] CephaNet on two different test datasets, for ranges of 2 mm, 2.5 mm, 3 mm, and 4 mm. For the first test dataset, the approach achieved 82.5%, 86.2%, 89.3% and 90.6%, while for the second test dataset, the results were 72.4%, 76.15%, 79.65%, and 85.9%. The study highlights the advantage of CephaNet over the original Faster R-CNN model in all ranges, with at least 10% better detection accuracy. In Chen et al. (2021) [58], Localization Accuracy was evaluated using a two-step process involving 3D Faster R-CNN for initial predictions followed by refinement with MS-UNet on the CBCT dataset. This combined approach resulted in improved accuracy, with initial measurements of 0.79 ± 0.62 mm and refined results of 0.89 ± 0.64 mm. Furthermore, on a diverse dataset characterized by a range of landmark counts, 3D Faster R-CNN alone achieved a localization accuracy of 2.34 ± 1.31 mm, highlighting the significant improvement achieved by subsequent MS-UNet refinement.

Dice was also mentioned in 4 studies. Le et al. (2017) [40] achieved a Dice coefficient of 0.83 for their ENet architecture. Goutham et al. (2019) [43] reported a dice of about 88% for each landmark. In Belikova et al. (2021) [56], their 3D U-Net obtained a dice of 0.682 ± 0.180. Finally, Tan et al. (2022) [80] achieved a dice coefficient of 54.25% for the proposed architecture.

Precision was referred in 3 studies. In Liu et al. (2020) [49], the precision of the model was reported to be 96.02%. Danilov et al. (2021) [59] achieved a precision of 0.97 for their ResNet V2 FT. Caspersen et al. (2022) [72] complemented their phase 1 architecture evaluation, reporting a precision of 90.6%.

Recall was reported for the same 3 studies. Liu et al. (2020) [49] obtained a recall of 92.86%. As above for precision, Danilov et al. (2021) [59] achieved a recall of 0.97 for their ResNet V2 FT, and Caspersen et al. (2022) [72] reported a recall of 89.7%.

F1 Score was mentioned in 2 of the above studies. In Liu et al. (2020) [49], the value was 94.68%. Danilov et al. (2021)

[59] reported a Macro F1 Score of 0.93 and a Micro F1 Score of 0.97.

Percentage of Correct Keypoints (PCK) follows with 2 studies. Tiulpin et al. (2019) [46] evaluated the Hourglass Network architecture with transfer learning techniques, comparing two different test datasets. The results presented concern the second stage of the architecture, related to the task of refining landmark detection. For ranges of 1 mm, 1.5 mm, 2 mm and 2.5 mm, the first test dataset obtained PCK values of 14.60 ± 4.83, 47.52 ± 2.20, 78.88 ± 0.88, and 93.48 ± 0.44, while the second test dataset achieved results of 11.24 ± 0.34, 44.98 ± 0.68, 75.12 ± 2.71, and 92.11 ± 0.34, respectively. The CNN proposed by Xu et al. (2019) [47] computed for the 5 mm and 10 mm thresholds, achieved PCK values of 77.8% and 96.4%, respectively.

Intraclass Correlation Coefficient was reported in 2 studies. Zhang et al. (2021b) [70] reported an overall kappa Intraclass Correlation Coefficient between two human raters and between one rater and the two-stage network model of over 0.9. Leitner et al. (2022) [77] evaluated their U-Net based architecture with an Intraclass Correlation Coefficient of 0.88.

The remaining classification metrics were only reported in 1 study each. The Error Detection Rate was used in Kang et al. (2021) [60] to evaluate the performance of an attention-modulated U-Net architecture. For the Digital Hand Atlas dataset, the hybrid perturbator approach produced an error detection rate of 4.76%, 0.96%, and 0.27% for the 2 mm, 4 mm, and 10 mm ranges, respectively. For the 2 mm, 2.5 mm, 3 mm and 4 mm intervals, using an edge detector perturbator, the subset Test 1 of ISBI2015 dataset produced an error detection rate of 13.26%, 7.89%, 4.91% and 1.79%, while the Test 2 subset showed results of 25.21%, 18.05%, 12.95%, and 6.79%.

Misdetection Rate is an error and failure rate mentioned in Lang et al. (2021) [62] to evaluate the two-stage DLLNet, achieving an impressive result of 0% (with unknown range threshold).

Success Rate was used in Tabata et al. (2021) [66] to evaluate their based Faster R-CNN architecture. With a range $R \in \{0, 2.5\}$ mm, the first test dataset achieved a success rate of 97.53%, while the second test dataset obtained 97.26%.

Precision rate was used in Torres et al. (2021) [67] to evaluate the VGG19-based CNN on two different datasets. For the synthetic dataset, the precision rate started at almost 85% for a distance error threshold of 0 mm and remained high, starting to decrease slightly after about 10 mm. For the real dataset, the precision rate started lower, just above 60%, improving rapidly until it stabilized at just over 90% for thresholds above 15 mm.

Identification Rate was used in Wang et al. (2021) [68]. For end-diastolic and end-systolic frames, the correct landmark identification rate was 83.3%.

Successful Classification Rate was reported in Kwon et al. (2021) [61] with average rates of 85.19% for the first test dataset and 81.96% for the second one.

Finally, the Failure Rate was used in Fard et al. (2021) [74] to evaluate their PoseNet architecture with L2 Loss. Failure rates of 2.77% for the normal subset, 3.33% for the extension subset, and 9.82% for the flexion subset were achieved.

3.3 Discussion

This systematic review provides an overview of the state of the art in DL methods for automatic anatomical landmarks detection. The head is the main body area analysed, with a total of 26 studies. This may be due to the availability of raw data and labelled datasets for the head, particularly for cephalometric analysis (13 studies), providing researchers with a ready-made resource for training and testing DL models. Challenges such as IEEE ISBI 2015 [96] also lead to increase data availability and the development of specific architectures for particular pathologies, which will then be used to benchmark the performance of future studies. This fosters a competitive and collaborative environment for DL advancements, meaning that more studies will

subsequently be found for the areas of the body included in the challenges.

In each clinical area, the number and variety of landmarks to be identified varies considerably between studies, even within the context of a common diagnosis. This variability may be due to the lack of standardised protocols that dictate the number and type of landmarks required, leaving researchers free to choose those they consider most appropriate. As a result, researchers may choose to combine established methods or develop new ones to optimise results. Such differences are further accentuated and incomparable between regions, given the distinct anatomical specificities and diagnostic requirements of each body region.

Among the imaging modalities reviewed, X-ray emerges as a predominant choice due to its wide availability and cost-effectiveness. The availability is simple and with fast accordance. Its low cost budget is combined with a major difference, since it is usually the first modality considered in the orthopedics area. These advantages, coupled with the capability to provide clear images of internal structures, considering bone analysis [97], make X-ray a preferred modality for tasks such as the automatic detection and image analysis. However this type of image has its limitations. Compared to other imaging modalities, it offers lower resolution and lacks spatial context. Soft tissues are not as clearly visualized with X-rays, which limits their diagnostic utility for certain areas of the body. For these reasons, CT scans are often the second choice, followed by MRI and US [98]. MRI and US are particularly advantageous for imaging soft tissues. US has the added benefit of being radiation-free, similar to MRI, but it is more cost-effective than MRI. In addition to the medical imaging modality, dataset size is also an important characteristic of the input data for DL models. However, assessing the impact of dataset size is complex due to the unique characteristics of each modality, such as resolution, format, and image size, which determine the volume of images required for optimal model performance. This complexity is highlighted when comparing 2D and 3D data. The latter introduces an additional dimension that increases both the input vector size and the computational requirements. Comparing to 2D dataset, 3D datasets, while potentially smaller in size (unit wise), can offer more comprehensive insights into the spatial relationships between anatomical structures due to their depth dimension, enhancing algorithm processing and consequently architecture performance [99]. It is safe to say the suitability of both 2D or 3D data will depend on several factors, the quality of imaging parameters, its curation process, provided information and the specific requirements of DL algorithm employed [100]. 3D images often need more sophisticated network architectures and a larger volume of training data to capture the intricate patterns inherent in the data effectively [39]. Studies with access to larger datasets, especially when matched in modality and research objectives, are likely to yield more robust findings, contingent also on the complexity of the DL model used [101].

Regarding data limitations, various solutions are usually applied to mitigate the problem. Data augmentation is one of these solutions. In Torres et al. (2021) [67] there are examples of both 2D and 3D data augmentation methods. The success of these studies, not only in the results but also in the data processing phase, could encourage the creation of similarly comprehensive and annotated datasets for other body areas, leading to a wider spread of research focus. Transfer learning is another well-established solution in the current context of image analysis, particularly in medical image [102]. A great percentage of studies have employed transfer learning, using VGG19 or ResNet as pre-trained models for initial mapping and feature extraction from images. The use of unsupervised learning could also be the future and one of the best methods to mitigate this data scarcity. One example is Generative Adversarial Networks (GAN), through the creation of synthetic datasets, as for example in Eslami et al. (2020) [48], which mentions the *pix2pix's* generator network. However, given the complexity of medical imaging, this method can present some problems. Image quality is a crucial factor due to its clinical implications. When generating data, researchers must consider the inherent complexity of each modality, including specific components and attributes such as dimensionality, resolution, contrast, and noise levels.

With regard to the DL architectures used, the diversity and complexity of the CNN architectures in the various studies is remarkable, highlighting the adaptability of these models to the respective tasks. A range of architectures, from U-Net and ResNet to more specialized networks such as Faster R-CNN, Hourglass Networks and DeepLabv3, demonstrate the rich set of tools and the availability of solutions on the field.

It is impossible to state that one architecture/model works better with a particular imaging modality than another, since each task is a unique challenge, dependent on several factors. The use of hierarchical models and multi-task learning approaches indicate a more controlled environment in which complementary systems are adaptable and capable of meeting the multifaceted challenges in medical imaging. However is possible to suggest trends that may emphasize better architectures than others. All the architectures analysed belong to a task where they fit and perform better. Segmentation, pose estimation, object detection and so on could be the subject of a more detailed and in-depth landmark detection challenge. The most widely used DL architecture in general was U-Net, addressed in 15 studies. It is used by at least one study concerning each of the identified body areas, with the exception of the knee, where no reference was made to the architecture. Furthermore, all medical imaging modalities, with the exception of the TAVI video imaging series, reference U-Net. Overall, the popularity of this architecture is well represented in the review, describing the broad applicability of the architecture, specifically its effective design for image segmentation tasks, since it comprises a contracting path to capture context and an expansive path for precise localization. Both ResNet (11 studies) [103] and VGG (3 studies) [104] have an easy access to be integrated with U-Net and other architectures, since it supports better gradient flow and training stability, thanks to ResNet's shortcut connections and VGG's effective depth.

The variability in architectures, the broad tasks in DL, and the specificity of the landmark detection task contribute to the diverse methodologies of pipelines identified in Table 3.2. While Direct Heatmap Probability Mapping is a commonly employed approach, it is often used in conjunction with other methodologies, such as Probabilistic and Regression approaches, highlighted in 7 studies. This combination does not necessarily imply that Direct Heatmap Probability Mapping is the primary pipeline. For example, in Ma et al. (2020) [50], this pipeline may be considered secondary as the study incorporates a Center of Mass (CoM) layer within a 3D U-Net-like architecture for interpreting network outputs, concerning heatmaps. Another type of methodology is the use of both Global to Local feature analysis and Direct Heatmap Probability Mapping. This combination is highlighted in 2 studies. In Zhang et al. (2021b) [70] initially, the study applies a shallow ResNet-50 network in a global manner to cover the entire image or volume, aiming to narrow down areas of interest based on overall image features that could be conceptually similar to heatmaps (in terms of highlighting areas of interest based on model predictions), the primary method does not involve creating and analyzing conventional heatmaps directly. Instead, it focuses more on refining the outputs from a global perspective down to local specifics. The Classification-based Candidate Generation approach is one of the more versatile methods since is used in conjunction with other methodologies. For instance, it might be combined with Direct Heatmap Probability Mapping for refining candidate locations or with Global to Local Feature Analysis to further narrow down and precisely identify landmark positions after the initial broad classification presented in Yang et al. (2015)[38] and Caspersen et al. (2022) [72]. Lastly and with 3 studies is the Iterative Patch-based Refinement pipeline. Li et al. (2018) [42] is a pratical example it utilizes an iterative approach to refine the localization of landmarks. It starts with an initial prediction, which is then iteratively refined by focusing on smaller patches around the predicted landmarks for increased accuracy. The method combines DL with traditional iterative refinement techniques, enhancing the precision of landmark detection by repeatedly adjusting the predictions based on localized patch analysis.

Analysing the results, all the studies used regression (error) metric (38 studies) and/or classification (38 studies), to assess

the model's performance. Both ME and MRE were the most mentioned regression metrics with 7 studies each, strictly following the way the metric is actually written in the study. It was possible to identify that some of the metrics might actually evaluate and comprehend the same objective with different nomenclatures, noticeable when revising the metrics results section. On a different note is worth mentioning Euclidean distance as the type of the measurement that most studies based its error measurements, i.e the geometric nature of most regression metrics, indicating the model's or framework performance is evaluated based on how far off the predictions are from their true locations in the dimensional space associated. In general, the most used metric (regression and classification wise) was SDR, which evaluates the accuracy precision of the model under some specific thresholds, in 10 studies 3.2.6. Of the 10 studies that use SDR, 9 are directly related to Cephalometric analysis and using the most common range of thresholds mentioned, 2.0, 2.5, 3.0 and 4.0 mm. Future research and studies should notice these values are extremely dependent on the clinical area that the study approaches, and since the majority are related to the head, for other area, with other metrics, the optimal ranges may be different for benchmarking and academic results comparison.

3.4 Conclusion

This review explored the DL strategies for landmark detection in medical imaging, shedding light on the intricacies of dataset characteristics, imaging modalities, architectural choices, and performance evaluation metrics. In conclusion and answering what factors should be considered for anatomical landmarks detection in medical images based on DL algorithms, the optimal DL solution for landmark detection transcends a one-size-fits-all approach. An important finding of this review is the critical role of data dimensionality and quality. While 3D data offers enriched spatial information crucial for nuanced landmark detection, it simultaneously demands sophisticated computational strategies and architectures to manage its inherent complexity. In this context, U-Net emerges as a versatile architecture, widely adopted across various body areas and imaging modalities for its robust segmentation capabilities and efficient localization precision under a variety of different datasets and dimension. Its prominence is particularly notable in cephalometric analysis within X-ray imaging, underscoring U-Net's adaptability to diverse application scenarios. Furthermore, the integration of foundational models like ResNet and VGG with U-Net highlights a trend towards leveraging deep feature extraction and transfer learning to enhance model performance, especially in scarce data and missing generalization environments. These architectures are directly related to the pipelines, revealing Direct Heatmap Probability Mapping as the most employed approach, often combined with methodologies like Probabilistic and Regression approaches, adapting CoM strategies. Metric analysis across the studies underscores the prevalence of ME and SDR as key benchmarks for evaluating model accuracy and reliability. These metrics, particularly SDR, offer a nuanced understanding of model performance under various thresholds, reflecting the precision required in different clinical diagnostics. This review not only describes the current landscape of DL applications in medical imaging but also sets the stage of trends for future research directions aimed at refining landmark detection methodologies for enhanced clinical outcomes.

Part I

Dissertation Core

Chapter 4

Solution Architecture

Establishing a well structured plan strategy is crucial prior to developing the solution. This chapter outlines the methodology and research plan strategy followed. It also sketches a solution pipeline strategy, fundamental to detail several important topics, from system requirements to final deployment. Through this steps a comprehensive and methodical approach is ensured.

4.1 Methodology and Research Strategy

The dissertation followed a research methodology based on CRISP-DM framework, a well-known approach in data mining, data analytics, and data science projects. This methodology guided the project's phases, ensuring a comprehensive and systematic approach to the task. An overview of how each phase was approached in the project is described. The methodology's base follows the six sequential phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment (Figure 4.1).

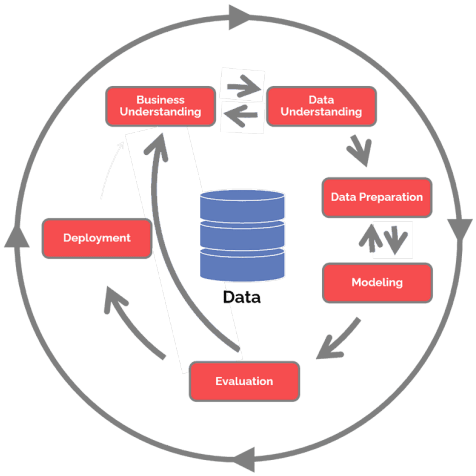


Figure 4.1: Research methodology based on CRISP-DM [105].

The framework guided the process of improving the accuracy and efficiency of PFI diagnosis using landmark detection in medical images, starting with an understanding of the task at hand, specifically its automatization. Under a specific context the Business Understanding stage was based on the interpretation of several factors crucial for a successful workflow and pipeline design. This involved analysing what indexes should be studied and consequently what landmarks were to be detected, correctly identify PFI risk factors. The decision was prior a research regarding information and insights of each patellofemoral

index. The knowledge was helpful to learn how to accurately mark and label the correct positions of these landmarks, a critical step. From the start, there were two possible approaches to the landmark detection task. The first approach considered was a 2D method, which involved identifying the optimum slice for each landmark and determining the position of each landmark within that slice. The second approach, and the one ultimately chosen, was a the identification of voxel position, in 3D method. The use of this approach, permitted the study of a less common scenario. By enabling the analysis of full spatial context on 3D data, the task converged to find the specific voxel corresponding to each landmark. Finally, besides the clinical part and approach identification, defining relevant metrics for evaluation was also important. All this retained information was obtained considering legal and ethical issues.

Data understanding was essential to, understand the anatomical structures in various imaging planes, assess quality and availability of data, and accuratly utilize the acquired datasets. First, in order to correctly perform image labelling a specific graphical interface was developed. This tool proved to be fundamental to a correct creation of the ground truth landmark positions and future steps. Through the annotation interface, image labelling was performed, storing every detail and important data regarding each volume images. This process was guided by a critical and extensive landmark annotation protocol.

Data preparation is the next phase, which involves the creation of a labeled dataset, also with the help of previous annotations. This is followed by a descriptive analysis to assure better execution of the preprocessing execution. The metadata obtained was crucial to the development of the ground truth dataset, which included removing outliers, creating correspondents heatmaps masks, resampling volumes to normalize scales, and serialization data for efficient modeling access.

The modelling phase involved developing models and variations of the same architecture, incorporating different DL aspects and technologies selected from an exhaustive study of the literature, as presented in Chapter 3. Based on the conclusions drawn, the architectures were derived from CNN, specifically utilising the 3D U-Net version [106]. This architecture was chosen not only for the good results from literature but also due to its proven effectiveness in medical image segmentation tasks, enabling precise voxel-wise predictions crucial for landmark detection. The architecture also influenced the resampling stage, ensuring that the input data was appropriately adjusted for fitting and training. The model specifications were fine-tuned according to system requirements, the dataset characteristics, and other criteria. Previous phases were fundamental for a correct pipeline follow-up.

Reaching to the evaluation step, this phase focused on the implementation of evaluation metrics that allowed to obtain the performance of the models. Finally, benchmarking was conducted to compare the selected architectures with existing literature solutions, specific from Chapter 3 study, highlighting their effectiveness and identifying areas for potential enhancement.

Following the sequential stages of CRISP-DM, the deployment stage aimed to integrate the best-performing model from the previous evaluation into *OrthoKnee*. However, this stage was not reached and will therefore not be discussed further.

4.2 System Requirements and Specifications

Prior to any kind of implementation, the development requirements were identified. Accordingly, the hardware and software were selected and their specifications were presented.

4.2.1 Hardware

The hardware used was determined by availability, always considering the requirements of the DL tasks to be performed. The GPU was chosen for its robust processing power, memory capacity and compatibility and compatibility with DL libraries

and computation. RAM was also a critical specification to consider. Since the dimensions of the images to be processed were 3D/4D, efficient memory management was essential. Handling the large data volumes and corresponding masks required ensuring they were readily available in cache during the development and tuning of the main models. On this note, the specific hardware used were two computer towers, with their specifications presented in Table 4.1.

Table 4.1: Hardware specifications

Device	OS	CPU	GPU	VRAM
Computer 1	Ubuntu 20.04	Intel Core i9-10940X CPU 3.30GHz	NVIDIA GeForce 3080 Ti	12 GB
Computer 2	Windows 10	AMD Ryzen 5 5600X 6-Core CPU 3.70GHz	NVIDIA GeForce 3090	24GB

Both computers were important to the process, but Computer 2 was preferred primarily due to its availability. The early stages of model design, training, and fine-tuning were conducted on Computer 1; however, its availability was limited compared to Computer 2. Ideally, it is not recommended to perform these procedures on GPUs with different specifications, especially when comparing DL model results. However, given the opportunity to conduct parallel work on two different computers, this approach was utilized.

4.2.2 Software

The software selection followed a specific configuration for each of two computers used in model training, influenced by the operative systems and their specifications. These specifications, in Table 4.2, were responsible for the downloading and installing the essential components and libraries. In addition to Python, CUDA and the correspondent cuDNN libraries were installed, with compatible and optimized version selected according to the needs. In addition, a miniconda environment was created to run the kernel with with the necessary Python packages. Setup ensured a robust and efficient environment for both operative systems. The development of the data labelling interface, ImageLabelGUI, followed the same principal of environment construction. This application was implemented in a normal *venv* environment.

Table 4.2: Software Specifications

Device	OS	Python version	CUDA version	cuDNN version	Tensorflow version
Computer 1	Ubuntu 20.04	Python 3.7.5	10.1	7.6.5	2.2.0
Computer 2	Windows 10	Python 3.8.18	11.2	8.1	2.10

4.3 System Pipeline

The stages of the CRISP-DM methodology framework guided the design of the system pipeline, shown in Figure 4.2.

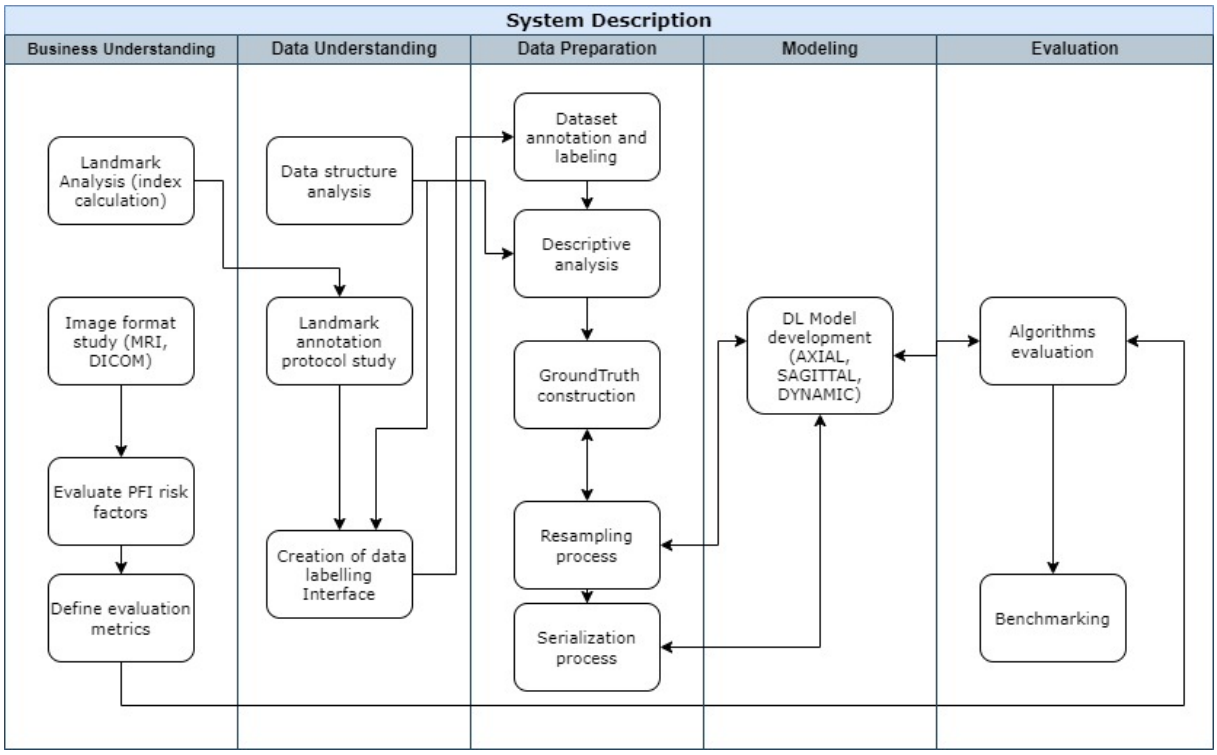


Figure 4.2: System pipeline under CRISP-DM guidelines.

4.4 Conclusion

The chapter explores a comprehensive framework of meticulous processes that were the basis for the research and development of the dissertation work. In terms of development requirements, both hardware and software were strategically chosen to meet the high demands of DL computations. The hardware setup with powerful GPUs and sufficient RAM allowed large datasets and complex computations to be handled efficiently, while the tailored software environment ensured compatibility and maximized performance.

From the initial understanding of business needs to the final evaluation and benchmarking, each phase was carefully planned. The diagram illustrates the comprehensive nature of the research strategy employed. Following the CRISP-DM guidelines, the dissertation work achieved a harmonious balance between theoretical research and practical application, leading to a set of solutions that are both scientifically valid and clinically relevant. This structured pipeline has facilitated the development of models that significantly enhance the accuracy and efficiency of PFI diagnosis, demonstrating the potential of advanced data analytics to improve medical diagnosis. It is worth to highlight the increased number of steps involved in data processing, which helps to reinforce the importance of data in tasks of this kind.

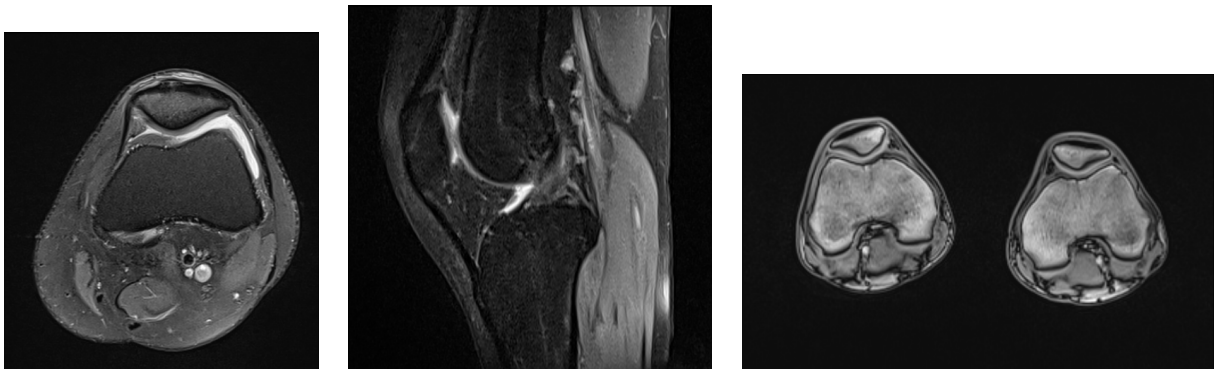
Chapter 5

Data Labeling

The system's development framework began with business and data understanding, crucial for all pipeline stages. This involved normalizing PFI evaluation indexes. Initially, the dataset was raw, requiring extensive filtering and labeling. To facilitate this, a GUI was developed to accurately mark landmarks in the image volumes, facilitating precise labeling and adjusting to the dataset's specifications.

5.1 Dataset structure

The raw dataset comprised knee MRI scans from 95 patients (1031 sequences), with a total of 140 knees (left and right), acquired at the *Hospital Trofa Saúde Braga Centro*. All data was fully anonymized to ensure that no identifiable information about the subjects was available, thus protecting their privacy and complying with all legal and ethical standards. Upon assessing the data, it was observed that the MRI volumes included different types of sequences according to the type of acquisition (Figure 5.1), specifically axial, sagittal and dynamic scans. The sequences were then separated according to the plane in which they were acquired, creating 3 different subsets of data: DATASET_AXIAL, DATASET_SAGITTAL and DATASET_DYNAMIC.



(a) (b) (c)

Figure 5.1: Type of acquisition planes: (a) axial, (b) sagittal and (c) dynamic.

Each of the subsets was organised by individuals, numbered from 1 to the total number of individuals in that subset, with information for the left, right, or both knees. Each sequence consists of a 3D volume composed of a set of 2D slices. Each sequence is stored in a DICOM file, an appropriate format commonly used in the medical imaging world. Each DICOM file contains not only the 2D images, but also a set of metadata related to the image acquisition method, which is essential for analysis, preprocessing and use of the data. Each subset required a cleaning, in order to successfully label the data.

Consequently each one was filtered according to the following exclusion criteria:

Criterion 1 : Knee anatomy hard to visualize;

Criterion 2 : Incomplete sequences (failure to acquire the entire knee);

Criterion 3 : Artefacts (e.g. screws from previous surgeries) that make it impossible to identify the desired anatomical points;

After filtering and splitting according to exclusion criteria, 38 sequences were removed from the raw dataset: 5 from DATASET_AXIAL, 5 from DATASET_SAGITTAL, and 28 from DATASET_DYNAMIC. With the dataset counting a total of 993 sequences across the three subsets.

Subset's MRI Data Types

An analysis was performed to understand the types of MRI sequences present in the different subsets of data, such as T1-weighted images, eTHRIVE for dynamic subsets, T2-weighted images, and Proton Density (PD) images, Table 5.1. The metadata presented was a DICOM analysis search in each subset's data.

Table 5.1: Distribution and total number of RMI sequences for the axial, sagittal, and dynamic data subsets

Subset	**T1-weighted**	**T2-weighted**	**PD**	**Not Mentioned**	**Total Count**
DATASET_AXIAL	83	0	151	0	234
DATASET_SAGITTAL	140	134	146	9	429
DATASET_DYNAMIC	330	0	0	0	330

For DATASET_AXIAL, the total number of sequences was 234, of which 83 were T1-weighted, 151 were PD, and none were T2-weighted. Regarding DATASET_SAGITTAL, the total number of sequences was 429. In the total subset, 140 T1-weighted sequences, 134 T2-weighted sequences and 146 PD/DP sequences were filtered. It was also be noted that 9 sagittal sequences were impossible to distinguish, where both the name and the DICOM data accessed were insufficient to assign the type of MRI. In the DATASET_DYNAMIC, the total number of sequences was 330, all of which were T1-weighted volumes.

Furthermore, the physical positioning of the knee was evaluated as well during each acquisition (e.g., in extension, flexion, in contraction, or without contraction), Table 5.2. This physical assessment was only applied to the DATASET_DYNAMIC. This analysis required the use of regex patterns to accurately categorize each subset's sequence. Only 1 sequence has no attributes (in extension, flexion, in contraction, or without contraction) justifying why the number of sequences with and without contraction total being 329.

Table 5.2: Distribution of dynamic MRI sequences by knee position and state of muscle contraction

Attribute	Without Contraction	With Contraction	**Total**
Flexion	74	74	148
Extension	79	77	156
Without flexion or extension	13	12	25
Total	166	163	329

After all this process of analysis and filtering, the data labeling phase started with an understanding of the protocol used to annotate the anatomical landmarks.

5.2 Anatomical landmark annotation protocol

Regarding the choice of indexes used to assess PFI, there is little consensus in the literature, and each hospital tends to develop its own protocol of indexes to use. However, a comprehensive review of Barbosa et al (2023) [11], part of the team, aimed to standardize this assessment by identifying the most crucial indexes for evaluating PFI. The review analyzed various protocols and incorporated the most relevant indexes. It was fundamental to enable the creation of a robust framework for a anatomical landmark annotation process.

The GUI, named ImageLabelGUI was designed to be comprehensive enough to cover the majority of existing protocols, reflecting the review's findings. This enabled anatomical markings (number of landmarks) to cover most of the existing indexes, and consequently most of the protocols. The review classified the predisposing risk factors into five groups: trochlear dysplasia (Figure 5.2), patellar height (Figure 5.3), patellar lateralization (Figure 5.4), patellar tilt (Figure 5.5), and tibial tubercle lateralization (Figure 5.6). The review identified key indexes that are critical for accurate PFI assessment. These indexes include:

Trochlear Dysplasia

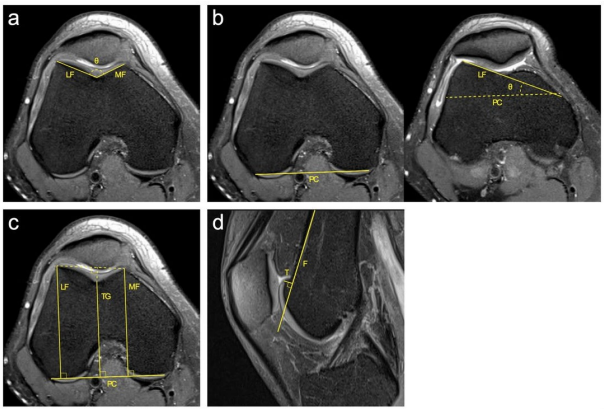


Figure 5.2: Measurements of the indexes to assess trochlear dysplasia [11]. (a) Sulcus Angle (SA) and Trochlear Facet Asymmetry, (b) Lateral Trochlear Inclination,(c) Trochlear Groove Depth, (d) Ventral Trochlear Prominence.

Patellar Height

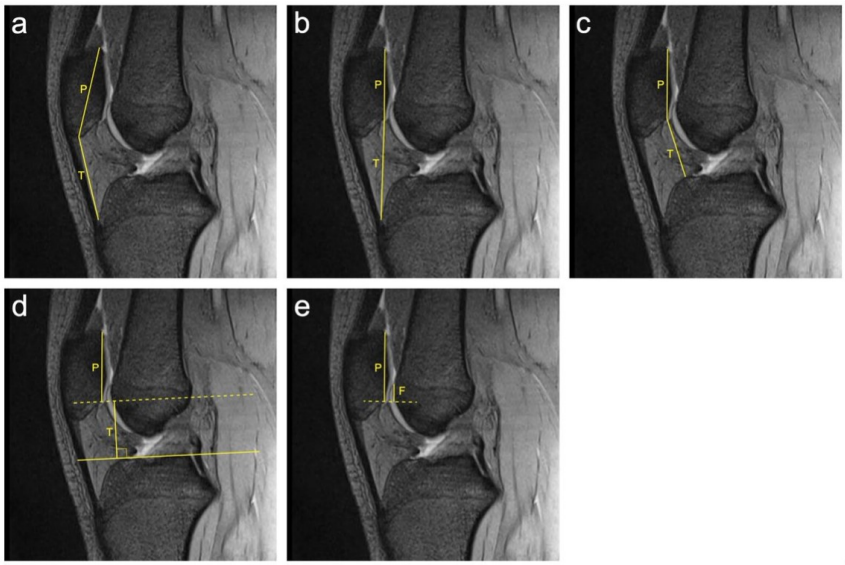


Figure 5.3: Measurements of the indexes to assess patellar height [11]. (a) Insall-Salvati Index, (b) Modified Insall-Salvati Index, (c) Caton-Deschamps Index, (d) Blackburn-Peel Index, (e) Patellotrochlear Index.

Patellar Lateralization

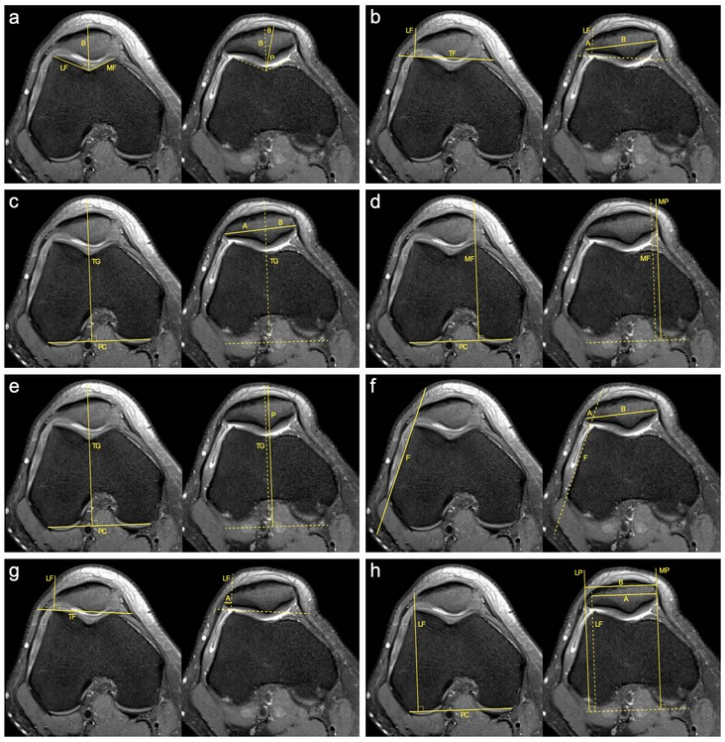


Figure 5.4: Measurements of the indexes to assess patellar lateralization [11]. (a) Congruence Angle, (b) Patella-Lateral Condyle and Lateral Shift, (c) Bisect Offset Ratio, (d) Laterall Patellar Displacement, (e) Patellar Displacement, (f) Lateral Patellofemoral Length and Tangent Offset, (g) Lateral Patellar Edge, (h) Patellofemoral Axial Engagement Index.

Patellar Tilt

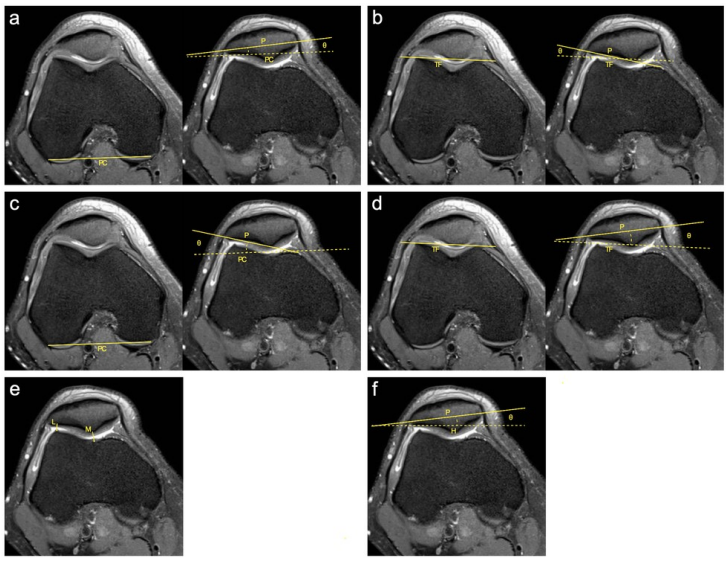


Figure 5.5: Measurements of the indexes to assess patellar tilt [11]. (a)] Patellar Tilt Angle, (b) Lateral Patellofemoral Angle, (c) Angle of Fulkerson, (d) Tilting Angle, (e) Patellofemoral Index, (f) Angle of Grelsamer.

Tibial Tubercle Lateralization

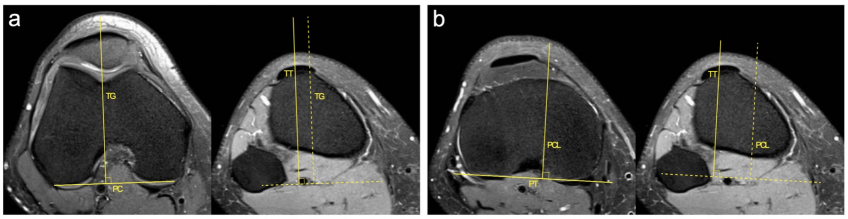


Figure 5.6: Measurements of the indexes to assess tibial tubercle lateralization [11]. (a) Tibial Tubercle to Trochlear Groove Distance, (b) Tibial Tubercle to Posterior Cruciate Ligament Distance.

Recognizing the need for standardization, the dataset was labeled following the comprehensive protocol mentioned above. This protocol includes indexes used by the Trofa Saúde group, ensuring that all marked landmarks support the assessment of these key indexes. The process required labeling points according to the protocol, with conditions like optimal slice detection and annotated points varying depending on the subset. Specifically, for each DATASET_AXIAL sequence, 11 anatomical landmarks were marked; for DATASET_SAGITTAL, 7 anatomical landmarks were marked; and for DATASET_DYNAMIC, a total of 18 anatomical landmarks were marked (Figure 5.7).

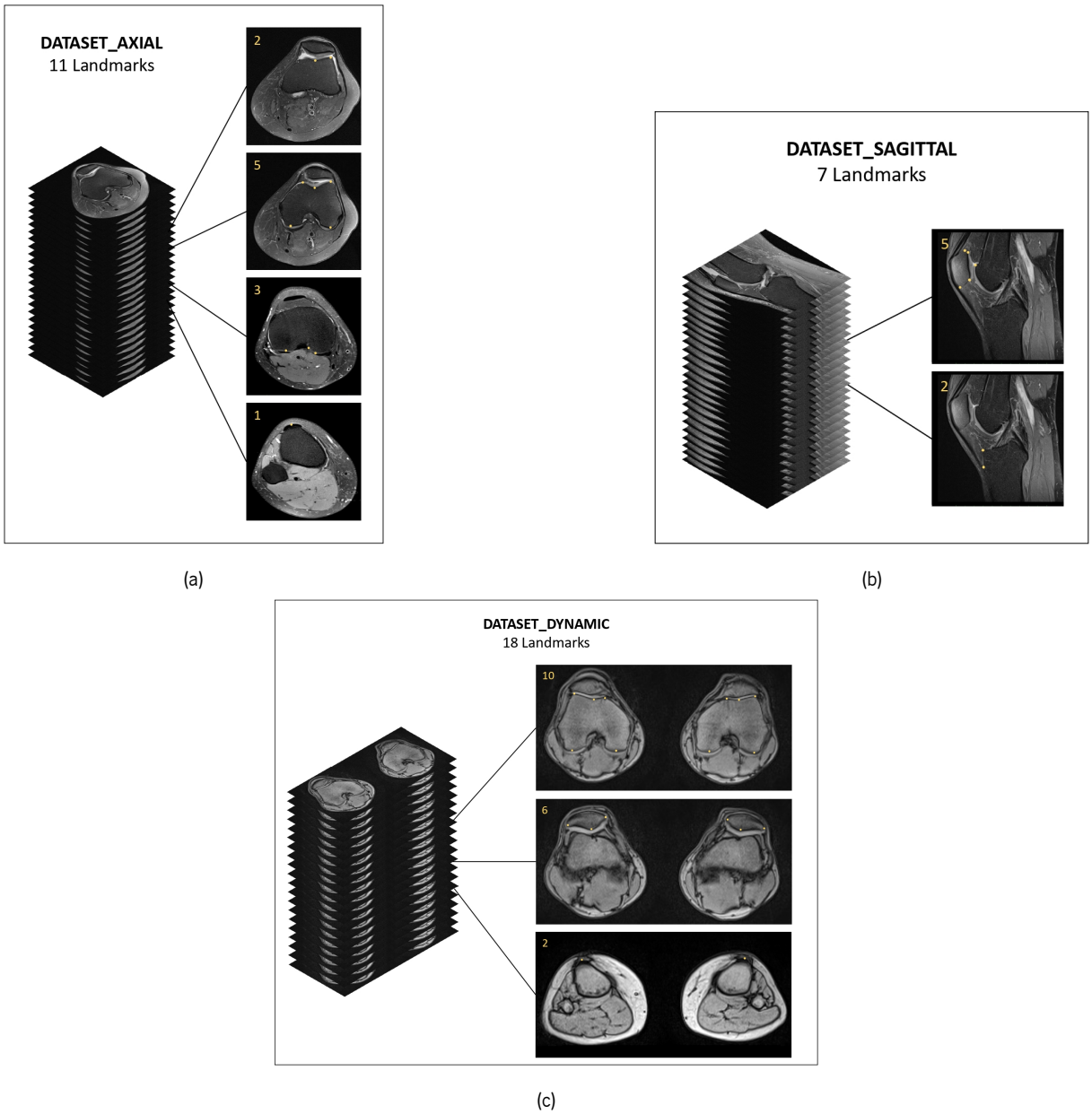


Figure 5.7: Number and location of anatomical points to label in the (a) axial, (b) sagittal, and (c) dynamic data subsets.

5.3 ImageLabelGUI

A specific tool, the ImageLabelGUI, was developed to label the anatomical landmarks in MRI images of the knee. This tool allows data visualization, image support for correct landmark labeling, annotation of anatomical landmarks, and status and storage of metadata. This platform allows the user to select a specific volume, navigate through its slices up to the maximum depth, annotate the maximum number of required or indicated points, and finally save these annotations.

The GUI was designed with a standard structure using Python 3.7.0. The toolkit incorporated a blend of PyQt5 and Visualization Toolkit (VTK) 8.1.2, along with Pandas, an xrld version of 1.2.0, and JavaScript Object Notation (JSON) packages.

PyQt5 was fundamental for composing the GUI interface, including the components that allow the application to be used. Components such as windows, buttons, text boxes, guidance labels and dock widgets were used to structure the interface. The

library uses layout managers to arrange these components in a structured way. PyQt5 also provides event handling capable of signal emission mechanisms required for specific actions when the components are used by the user. The integration of VTK with PyQt5 made it possible to use of this renderer to visualize medical images. For this integration, *QVTKRenderWindowInteractor*, a PyQt5 widget, embeds the VTK rendering into the Qt framework.

VTK is a C++ library accessible through Python wrappers. As an open-source software system specialized in image processing, 3D graphics, volume rendering, and visualization, VTK is well-suited for the task. The choice of this framework was primarily based on the need to render medical images. The manipulation of DICOM files was implemented using methods capable of parsing directories, reading pixel data, and applying modality and VOI LUTs for proper visualization of the volume rendering. The 2D slices of 3D volumes were navigated and displayed using VTK's *vtkImageReslice*, updating the displayed image and landmarks according to the selected sequence. Second, the features of personalized interactivity and image display customization were made possible by the functionalities of the VTK rendering window. In addition to the ability to adjust image properties such as window level and contrast, which are essential for medical image analysis and better landmark annotation, personalized interactivity was enabled by the *vtk.vtkInteractorStyleUser* class, which was to provide custom responses to mouse events such as panning, zooming, and clicking within the VTK window. These features proved to be very useful in the process of annotating the landmarks. Since VTK is a rendering application, these features dependent on the correct use of VTK camera and the *vtk.vtkCellPicker* picking mechanism. The latter enabled a dynamic landmark annotation function, allowing precise selection of points within the render.

5.3.1 Landmark Annotation process

ImageLabelGUI precise development was crucial for the correct landmark annotations. First, when chosen a sequence/volume to work on, the tool would check its state and provide monitoring while the process of marking points occurred. The process helped the user, i.e. a help image guide was implemented to lead the user through the process of marking points. When marking a point, the user was guided to the next landmark, which also helped identify the correct slice. This guide distinguished between left and right knees. Each subset had specific instructions for the annotations, detailed below:

Axial

The first subset to be labelled was DATASET_AXIAL (Figure 5.8). The annotation processed required finding the optimal slice for each set of landmarks (Figure 5.7a).

- **2 landmarks set**: Identify the optimal slice by locating the first image with visible trochlear cartilage (significant grey area). If uncertain, choose the slice with the more relevant cartilage (white stroke).
  - 1 landmark, mark on the lateral side, where the cartilage ends.
  - 1 landmark, mark on the lowest point of the cartilage (middle of the femur).
- **5 landmarks set**: Identify the slice between the femoral condyles, the deepest aspect of the intercondylar groove (appearance of a perfect Roman arch). For a tie-breaker select the slice with the greatest distance between the epicondyles (they resemble lateral beaks):
  - 2 landmarks on the condyles, marked at the lowest point of each curvature.

  - 3 landmarks, 1 on the lateral side at the highest point (not the end of the cartilage), 1 on the lower curvature at the lowest point, 1 on the medial side at the highest point (not the end of the cartilage).

- **2 landmarks set**: Locate the fibula (maximum visualization), then proceed until it disappears completely and the tibial plate is defined (containing the dorsal condylar line of the tibia, just below the articular surface of the tibia plateu, and above the fibular head).
  - 2 landmarks, first from bottom to top as soon as the fibula is no longer visible, mark the femoral condyles (as reference, draw a tangential line to the posterior aspect of the condylar line of proximal tibia).

- **1 landmark set**: Locate the fibula (maximum visualization), then proceed until it disappears completely where there is a more marked division between the cruciate ligaments and the tibial plateau.
  - 1 landmark, mark on the most medial side (differs from left to right knee), in the posterior cruciate ligament.

- **1 landmark set**: Locate the patellar tendon and follow it to the first image where it is fully inserted in the tibia tubercle (bone).
  - 1 landmark, mark the most superficial curvature (on the surface of the outer ligament) at the most central point.

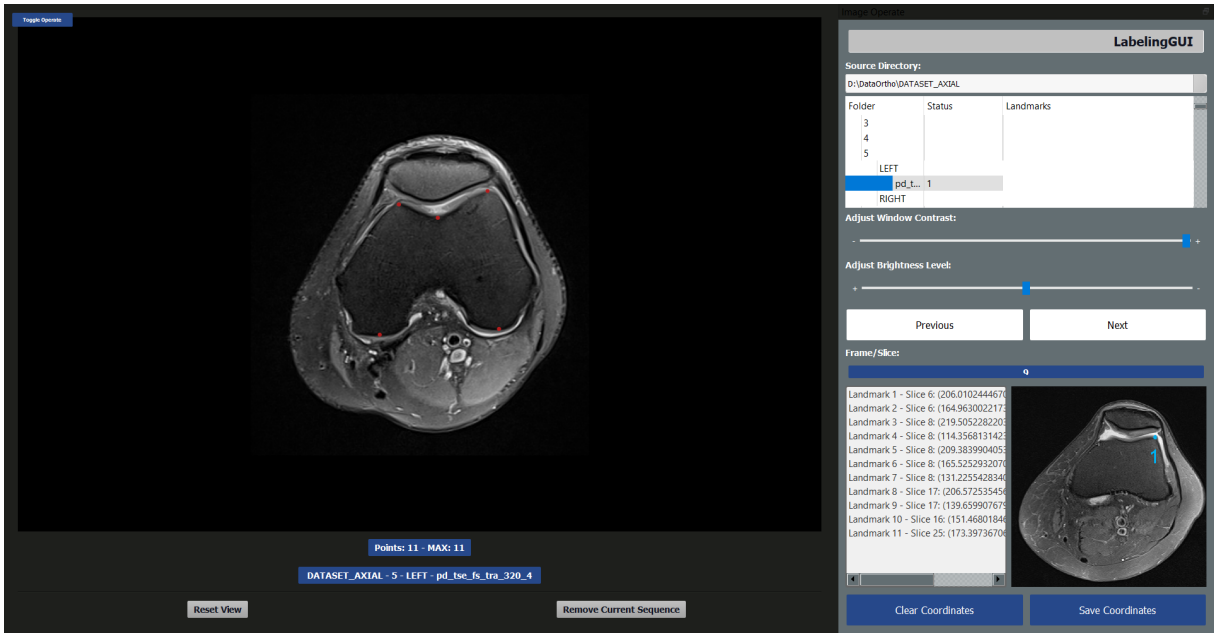


Figure 5.8: ImageLabelGUI with an example of landmarks annotation on axial slices.

Sagittal

Next was DATASET_SAGITTAL (Figure 5.9). Just like for the Axial subset, the annotation required finding the optimal slice for each set of landmarks (Figure 5.7b). In this case for each set of landmarks there could be a maximum of 3 optimal slices.

- **4 landmarks set**: Decision factor is the patella. Choose the slice with the longest axis of the patella.
  - 2 landmarks. Mark the top and bottom of the patella (at the beaks of the bone). Use the most proximal and most distal points. Draw the longest axis line between these two points.

- 2 landmarks On the axis of the cartilage (mark at the highest and lowest point).

- **1 landmark set**: Locate the slice where the cartilage is as close as possible to the femur. This landmark may not be on the same slice as the previous set.
  - 1 landmark, highest point of the femur cartilage (adjacent to the patella).

- **1 landmark set**: Identify the tibial tubercle with higher tuberosity (a large prominence on a bone usually serving for the attachment of muscles or ligaments) and the fusion of the patellar tendon with the bone.
  - 1 landmark. Mark exactly at the fusion point between the tendon and the bone, which is the largest portion of the tendon attached to the tibia.

- **1 landmark set**: Do not follow the patella. Choose the slice where it is visible the most proximal (top) and anterior (forward on the bone) part of the tibia.
  - 1 landmark, mark the point on the tibia that is most anterior on the horizontal axis and most proximal on the vertical axis, closest to the patella (see Canton-Deschamps point).

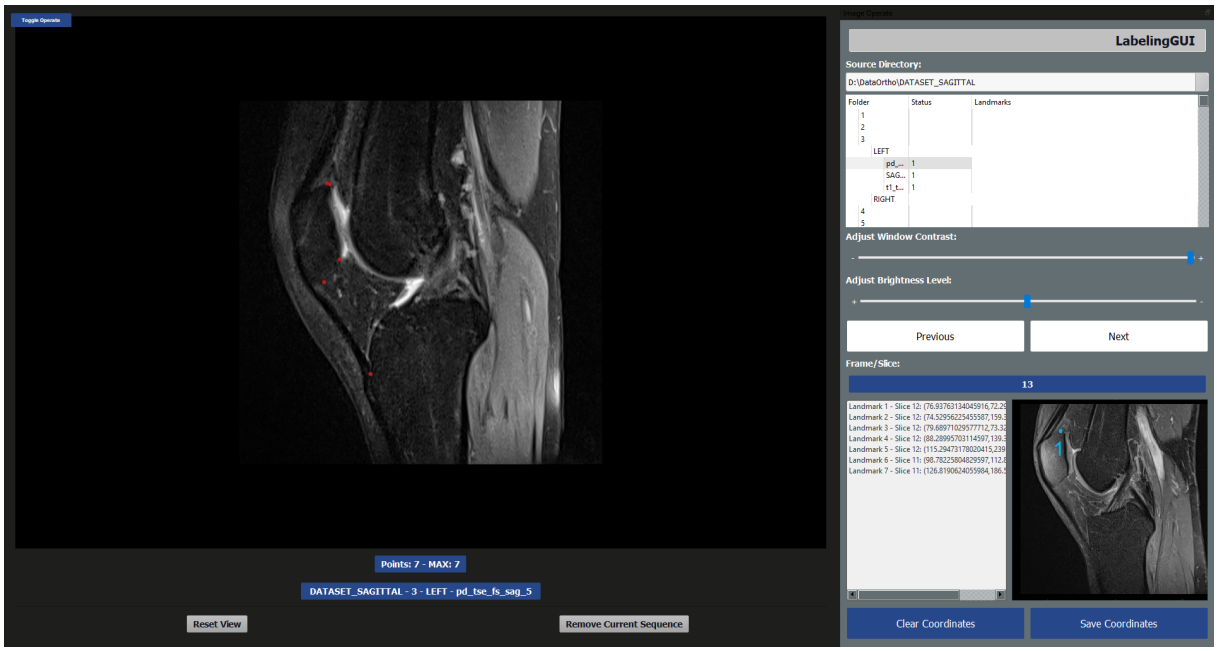


Figure 5.9: ImageLabelGUI with an example of landmarks annotation on sagittal slices.

Dynamic

Lastly was DATASET_DYNAMIC (Figure 5.10). In this subset given that this type of aquisition had both knees, the annotation was made on both knees, if 5 landmarks were marked on the right knee, the same were noted on the left, making a total of 10 landmarks. In this process it was fundamental to start the annotation from the right knee to the left (Figure 5.7c).

- **10** (2×5) **landmarks set**: Identify the slice between the femoral condyles, the deepest aspect of the intercondylar groove (appearance of a perfect Roman arch). For a tie-breaker, check the epicondyles (they resemble lateral beaks). Find a compromise between the two crowns on single optimal slice.

- 2 landmarks on the condyles, marked at the lowest point of each curvature.
- 3 landmarks, 1 on the lateral side at the highest point (not the end of the cartilage), 1 on the lower curvature at the lowest point, 1 on the medial side at the highest point (not the end of the cartilage).

- **6** (2×3) **landmarks set**: Identify the longest axis of the patella between the medial and lateral beaks.
  - 3 landmarks, in the patella, 1 lateral point, 1 superior point and 1 medial point. Each set of 3 landmarks could be in a different optimal slice.

- **2** (2×1) **landmarks set**: Locate the patellar tendon and follow it to the first image where it is fully inserted in the tibia tubercle (bone).
  - 1 landmark, mark the most superficial curvature (on the surface of the outer ligament) at the most central point.

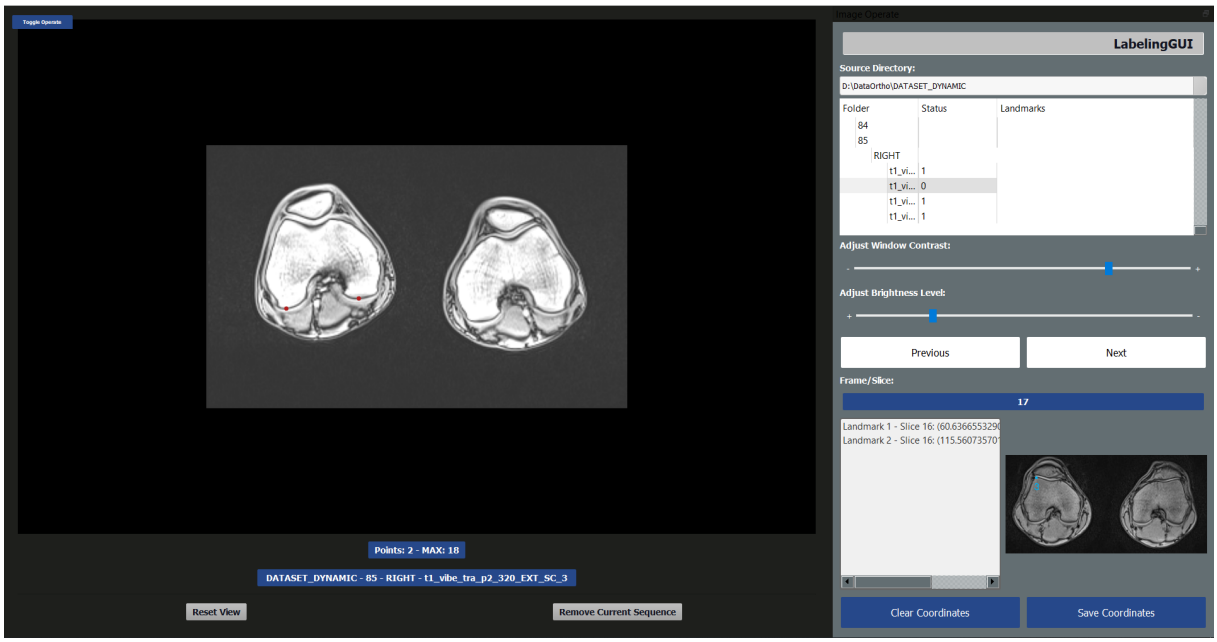


Figure 5.10: ImageLabelGUI with an example of landmarks annotation on dynamic slices.

After the process of visualizing and annotating the landmarks, it was necessary and crucial to define a state of the completed set of landmarks for each sequence of each subset and provide storage of it. The GUI's design took into account clinical workflows, such as the distinction between left and right knees, the different number of landmarks required for each subset, and their correspondent order. Providing a different status for each sequence, in completed and finalized.

The system also provided functionalities for saving the annotated landmarks, depending on whether the requirements defined for saving the landmarks were met. The anatomical landmark and associated metadata were saved in JSON and Excel formats to facilitate data management and future analysis for preprocessing and future modeling.

5.4 Conclusion

The data labeling process was foundational to the success of the entire project. Starting from a comprehensive understanding of the business and data requirements, the system's framework was meticulously designed to standardise and streamline the evaluation of PFI. The raw dataset, comprising knee MRI scans from 95 patients, was systematically organised to ensure consistency. Through rigorous filtering based on predefined criteria, the dataset was refined to exclude any sequences that could compromise the quality of the subsequent analysis.

The anatomical landmark annotation protocol, grounded in a thorough review of existing indexes, aimed to cover the majority of protocols used in clinical settings. This standardization was fundamental in ensuring that the annotations would be broadly applicable and clinically relevant. The developed GUI enabled users to select a dataset, navigate through its slices, and mark anatomical landmarks on the images. The interoperability between the different libraries allowed a correct workflow for the needs of the task, i.e for the correct landmarks annotation and its storage.

Overall, the chapter detailed a robust methodology for data labeling, emphasizing the integration of advanced software tools and standardised protocols to achieve high-quality annotations. This meticulous approach laid a solid foundation for the subsequent stages of the research, ensuring that the data used for model training and evaluation was of the highest possible standard.

Chapter 6

System Description

In various data science fields, particularly within analysis and processing domains, a thorough understanding and visualization of available data are crucial. The initial phase of our approach involved a comprehensive descriptive analysis of the dataset, focusing on data distribution, pattern identification, and data quality assessment. This stage also included the examination of associated metadata, which informed subsequent preprocessing decisions.

Specifically, in the context of MRI data, ensuring that the input data for models is clean, consistent, and accurately represents the biological structures under study is vital. The preprocessing phase, developed in Computer 1 addressed these needs by refining the subset data. High-quality data is paramount in supervised analysis problems, where the protocol significantly influences evaluation outcomes. The process emphasized the correct creation of ground truth, involving precise landmark annotation to enhance data reliability and model performance.

The rest of the chapter details the preprocessing and modeling steps crucial for handling MRI data. Initially, the ground truth for all data subsets, resampling MRI volumes and masks to ensure dimensional consistency and the transition to TFRecords, particularly using Computer 2, for efficient data handling and normalization processes to maintain consistent input scales. Subsequently, data augmentation techniques, such as random rotations, flips, translations, blurs, and noise injections, were employed to increase dataset diversity and model robustness. Finally, advanced DL architectures, including various forms of 3D U-Nets with residual and attention mechanisms, were implemented. These models utilized sophisticated convolutional layers, normalization techniques, loss functions, and optimizers, supported by callbacks to monitor and optimize the training process.

6.1 Descriptive Analysis

The role of descriptive analyses is fundamental to understanding MRI data. A comprehensive analysis was performed for each subset of axial, sagittal and dynamic data. During this phase, not only essential metadata, such as the number of slices (volume depth), slice thickness, and spacing between slices were capture and examined. For each subset, enhanced data was shown in terms of structure, quality, and metadata distribution, setting the stage for a customized preprocessing phase. For each subset, enhanced data was presented in terms of structure, quality, and metadata distribution, setting the stage for a customised preprocessing phase. The implementation of metadata graphs allows for a comprehensive visualisation of the distribution across different fields.

Figure 6.1 shows the distribution of slices by subset. When comparing within subsets, the DATASET_DYNAMIC subset is the most unbalanced in terms of the number of slices per volume. Notably, DATASET_DYNAMIC is the most unbalanced, with slice numbers ranging broadly from 30 to 90 slices per volume, showing two main peaks around 70 and 80 slices. In contrast, DATASET_AXIAL and DATASET_SAGITTAL display more concentrated distributions, with DATASET_AXIAL clustering around 30

slices and DATASET_SAGITTAL around 20 to 30 slices, peaking near 25 slices.

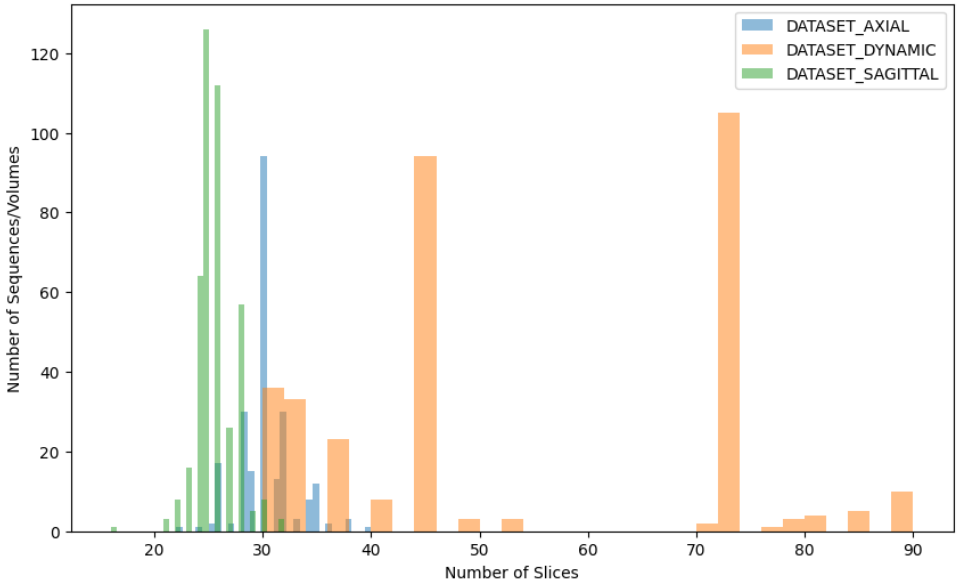


Figure 6.1: Distribution of the number of slices (volume depth) for each subset.

On another analysis, Figure 6.2 provides a different visualization, presenting the average of the previous shown dispersion. The volume depth represents a critical value for the data input and model's dataset requirements, providing a better perspective on the optimal values that could be used for data shape input for the models. The image highlights that DATASET_DYNAMIC has the highest average at approximately 52.63 slices, suggesting more detailed volumetric data. DATASET_AXIAL has an average of around 30.2 slices, balancing detail and processing load, while DATASET_SAGITTAL has the lowest average at approximately 25.64 slices. This information translates into an interpretation of the optimum values for this z-axis depth for each subset.

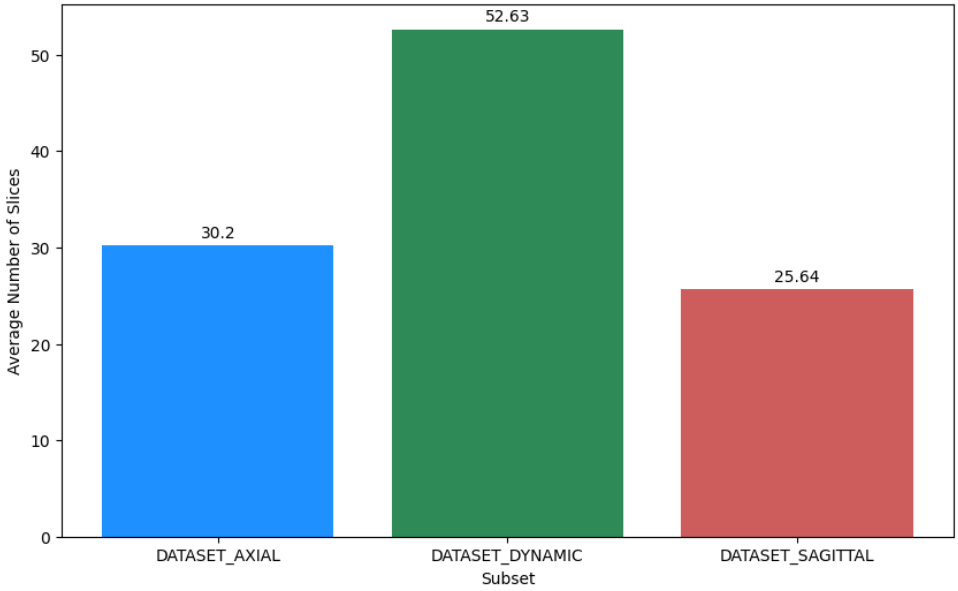


Figure 6.2: Average number of slices for each subset.

The average values of the rows and columns of the volume slices help to set the desired input at which the data should

be processed. This Figure 6.3 analysis illustrates these average values for each subset. It reveals significant differences that impact further preprocessing and modeling stages. For instance, the DATASET_AXIAL and DATASET_SAGITTAL subsets exhibit similar dimensions, with averages around 360mm for rows and slightly lower for columns, almost squared dimension slices. In contrast, the DATASET_DYNAMIC subset shows notably smaller average dimensions, particularly in rows, averaging around 243.85 mm. These variations are crucial as they influence the resolution and scaling requirements. These dimensions give insightful perspectives according to volume resampling, i.e future shape preprocessing could be more or less aggressive towards these rows and columns values, losing more or less pixel information. Ensuring that the input dimensions are standardized across different subsets will facilitate a more efficient and accurate training process, accommodating the unique characteristics of each data subset.

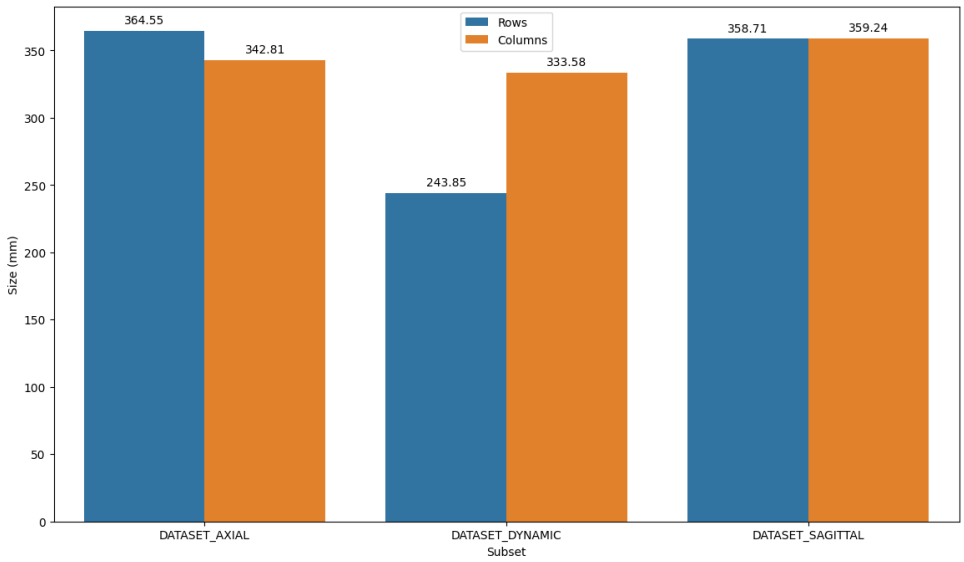


Figure 6.3: Average values of rows and columns of the slices for each subset.

The average pixel spacing values in the X and Y directions for each subset provide critical insights into the spatial resolution of the images, essential for accurate 3D reconstruction. Figure 6.4 illustrates these mean values.

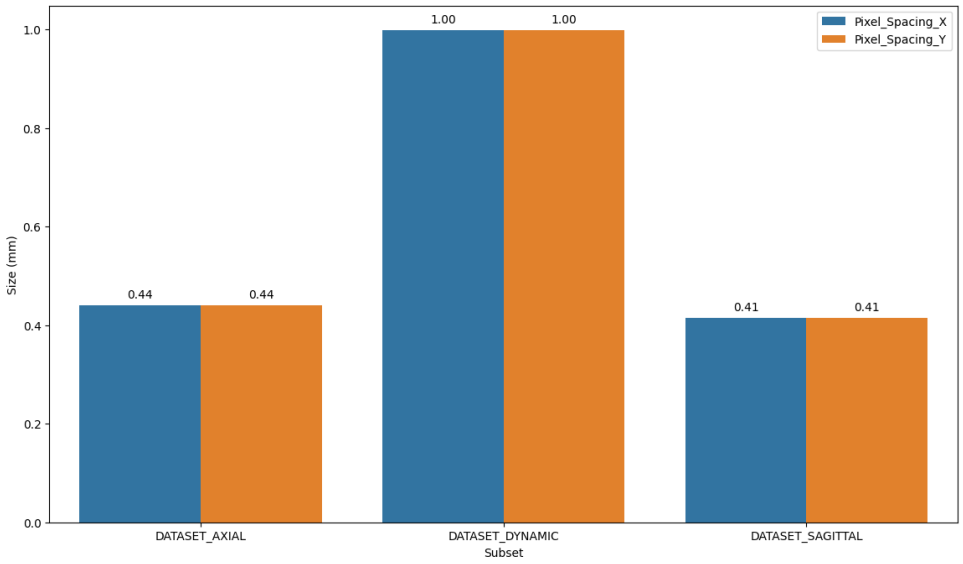


Figure 6.4: Average values of pixel spacing for the x and y axes for each subset.

In terms of pixel spacing for both x and y coordinates, all subsets had the same values, which simplified the processing and analysis by ensuring isotropic resolution. This uniformity in pixel spacing allowed for consistent application of image processing techniques and model training, reducing potential artifacts and ensuring accurate spatial measurements.

6.1.1 Individual Subset Analysis

A more detailed descriptive analysis was also carried out for each subset, with graphs visualization and subsequent analysis. For each of them, a study of the distribution of the metadata distribution was conducted, specifically the following fields: a) slice thickness, b) spacing between slices, c) number of slices, d) rows, e) columns, and f) pixel spacing. The study of these fields was important because it aimed to study the influence of 3D input data on DL algorithms.

Axial Subset

A more detailed analysis of distribution to the DATASET_AXIAL is provided in Figure 6.5. The distribution analysis is made for slice thickness, spacing between slices, number of slices, rows, columns, and pixel spacing.

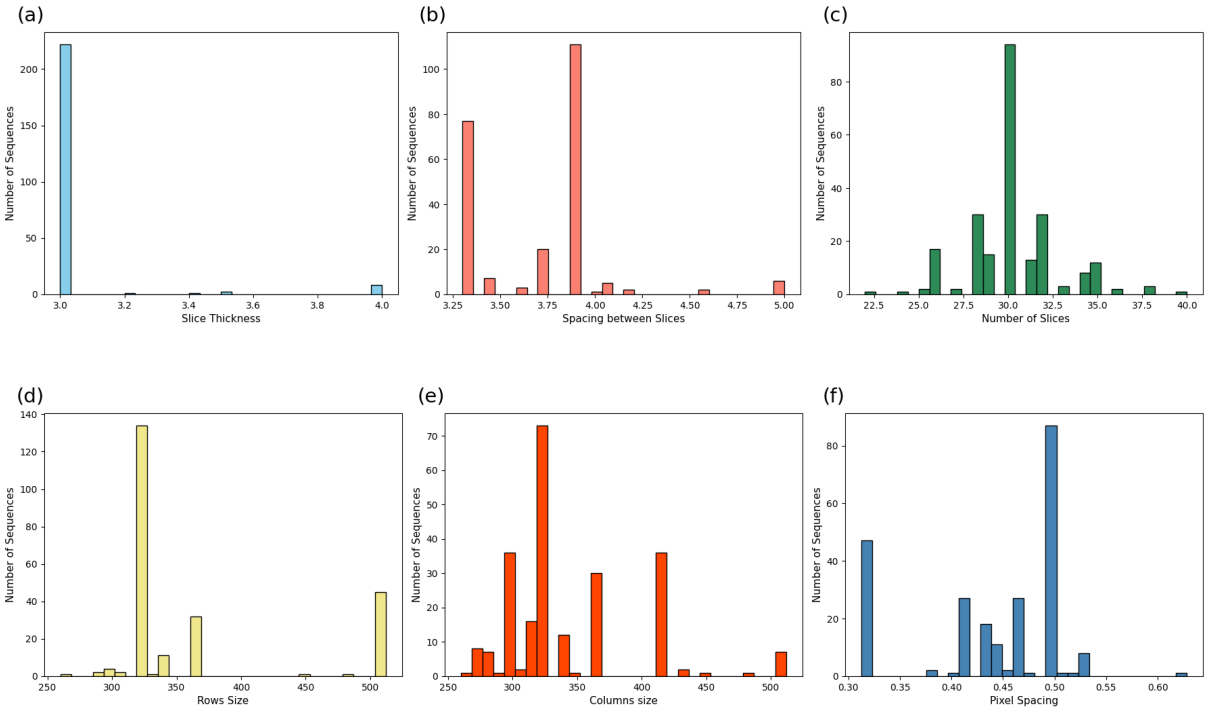


Figure 6.5: Distribution fields for DATASET_AXIAL.

Looking at the distribution of the slice thickness parameter, the histogram shows very consistent values, mainly around 3.0 mm, with almost no variation, suggesting a highly standardized protocol for slice thickness in this axial subset. The spacing between slices shows two distinct peaks, one at 3.3 mm and the other at 3.9 mm. This suggests that there were two predominant settings or protocols axial slices spacing in the scans. Regarding the number of slices per volume, there was a high concentration around 30 slices, with equivalent margin of error for values above or below this mark. The rows histogram, which corresponds to the vertical size of the images, shows that the majority of the images have a vertical size of 320 pixels, with a minority having a vertical size of 512 pixels. The columns histogram showed a varied distribution with three notable peaks around 300, 320, and 416 pixels, suggesting a variety of image widths. Pixel spacing values showed a significant peak at 0.5 mm, with a smaller peak at 0.31 mm. This suggests perhaps, that two different pixel densities were primarily used in the axial image scans, resulting in different image resolutions.

Sagittal Subset

The same procedure analysis of distribution was made to the DATASET_SAGITTAL, in Figure 6.6. The distribution analysis is made for slice thickness, spacing between slices, number of slices, rows, columns, and pixel spacing.

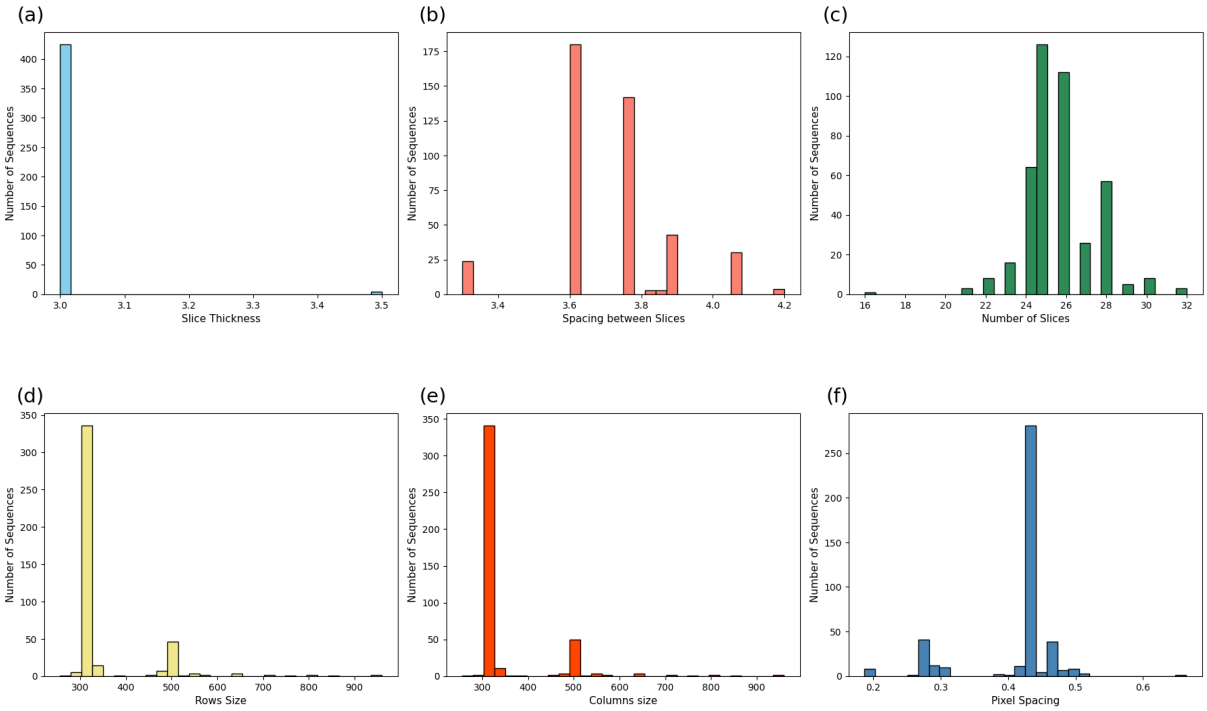


Figure 6.6: Distribution fields for DATASET_SAGITTAL.

Assessing the different sagittal subset data, the slice thickness histogram showed that the values were very consistent, with the majority of the slices being 3.0 mm thick. Only an almost insignificant number of slices had a thickness of 3.5 mm. The spacing between slices showed two main groups, around 3.6 mm and 3.75 mm. This suggested that there were two distinct protocols or scanner settings used in the dataset with respect to the distance between slices. The number of slices per volume seemed to be variable, with peaks at around 24, 25, and 26 slices. The number of rows for the images in the sagittal dataset mainly centered around 320 pixels, with a considerably reduced amount around 512 pixels, suggesting a balanced image height across the subset. Similar to the distribution of rows, the number of columns was also predominantly around 320 pixels, with a minority around 512 pixels, indicating a uniform image width and, together with rows, suggesting that the images were likely to be square. The pixel spacing was predominantly around 0.44 mm, indicating that the resolution of the pixels in the image was consistent and again pointing to a standardized acquisition process.

Dynamic Subset

As in the previous subsets, Figure 6.7 illustrates the distribution of parameters within the dynamic subset, including slice thickness, spacing between slices, number of slices, rows, columns, and pixel spacing.

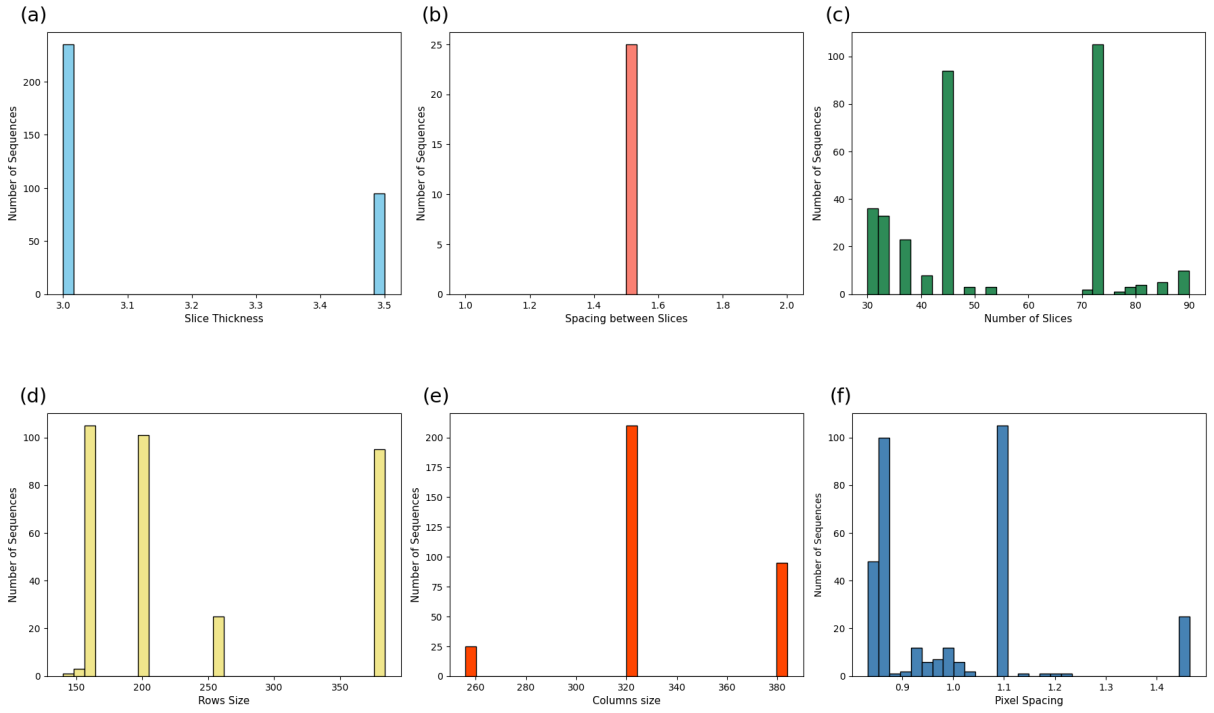


Figure 6.7: Distribution fields for DATASET_DYNAMIC.

Examining the metadata in the dynamic subset, the slice thickness histogram showed a highly concentrated number of occurrences around 3.0 mm, indicating a good balance across the dataset. A smaller number of volumes had a slice thickness of 3.5 mm. The only value for the spacing between slices was 1.5 mm. The number of slices in the dataset varied, with peaks around 44 and 72 slices, indicating an unbalanced in the dataset with respect to this characteristic. The histogram of rows showed three prominent peaks of approximately 160, 200, and 384 rows. The distribution of the columns was different from that of the rows, with three peaks around 256, 320, and 384, the last two values being considerably higher, confirming the fact that the 2D DICOM images for the dynamic dataset were essentially rectangular. The pixel spacing showed a distribution around 0.80 mm and 1.5 mm, with peaks at 0.86 mm and 1.46 mm. This variation suggested the presence of unbalanced pixel resolutions in the dynamic dataset.

6.2 Preprocessing

6.2.1 Ground Truth Construction

The process of creating ground truth for all data subsets was the first step in data preprocessing. Among the two approaches considered for ground truth generation—3D binary heatmaps and 3D Gaussian heatmaps—the latter was chosen for its capability in providing three dimensional information, helping the network in converging to the optimal point for each landmark's position. Gaussian masks offered a probabilistic representation that peaks at the landmark point and gradually decreases, enabling the network to better understand the proximity of voxels to the landmark. This probabilistic gradient facilitates smoother and more precise gradient updates during training. The primary purpose of these masks was to transform the task into a heatmap regression problem. On a first stage a mask was created for each anatomical landmark and for the background of each

volume, so that each 3D volume/sequence is associated with as many masks as there are landmarks marked for that subset plus a background. Thus, 12 masks were created for DATASET_AXIAL, 8 masks for DATASET_SAGITTAL and 19 masks for DATASET_DYNAMIC. A Sigma percentage was assigned to determine and control the spread and size of the Gaussian masks in the heatmap.

By generating these probabilistic masks, the networks were able to produce probabilistic heatmap areas indicating where the landmarks were likely to be located. The Equation 6.1, G_L, defines the Gaussian heatmap mask G centered on the landmark L. The value of the heatmap for a voxel (x, y, z) covers the interval [0, 1], where (x_0, y_0, z_0) is the center of the Gaussian and $\sigma_x, \sigma_y, \sigma_z$ are the standard deviations in the x, y and z directions, respectively. This probability distribution is defined by the distance from a voxel x, y, z to the coordinates of the landmark x_0, y_0, z_0. The goal was to approximate the behaviour of a spherical mask within the channel heatmap.

$$G_L(x, y, z) = \exp\left(-\left(\frac{(x - x_0)^2}{2\sigma_x^2} + \frac{(y - y_0)^2}{2\sigma_y^2} + \frac{(z - z_0)^2}{2\sigma_z^2}\right)\right) \tag{6.1}$$

The percentage function shown in Equation 6.2 was constructed to calculate the standard deviations of the Gaussian distribution with respect to the dimensions of the MRI volume, $Rows \times Columns \times Number\_of\_Slices$. The *percentageSigma* value is used to determine the standard deviation as a percentage of the respective dimensions. These values help in controlling the spread of the Gaussian distribution along each axis. To do this, a static value was chosen to apply for each standard deviation of the landmark's coordinate.

$$\begin{aligned}
\sigma_x &= \text{percentage}\left(\frac{\text{percentageSigma} \times \text{Rows}}{100}\right) \\
\sigma_y &= \text{percentage}\left(\frac{\text{percentageSigma} \times \text{Columns}}{100}\right) \\
\sigma_z &= \text{percentage}\left(\frac{\text{percentageSigma} \times \text{Number_of_Slices}}{100}\right)
\end{aligned} \tag{6.2}$$

The creation of ground-truth channel heatmaps involved repeating the Gaussian process, Equation 6.1, for each sequence volume in all subsets. The images in Figure 6.8 show the 3D Gaussian mask for the first anatomical landmark (landmark 0) for each of the axial, sagittal, and dynamic subsets. The masks show a concentrated bright spot that diffuses as the slice is moved, indicating the spread of the landmark throughout the volume. This consistent spread pattern across the different data subsets highlights the 3D positioning of the landmark and the diffusion of the mask area.

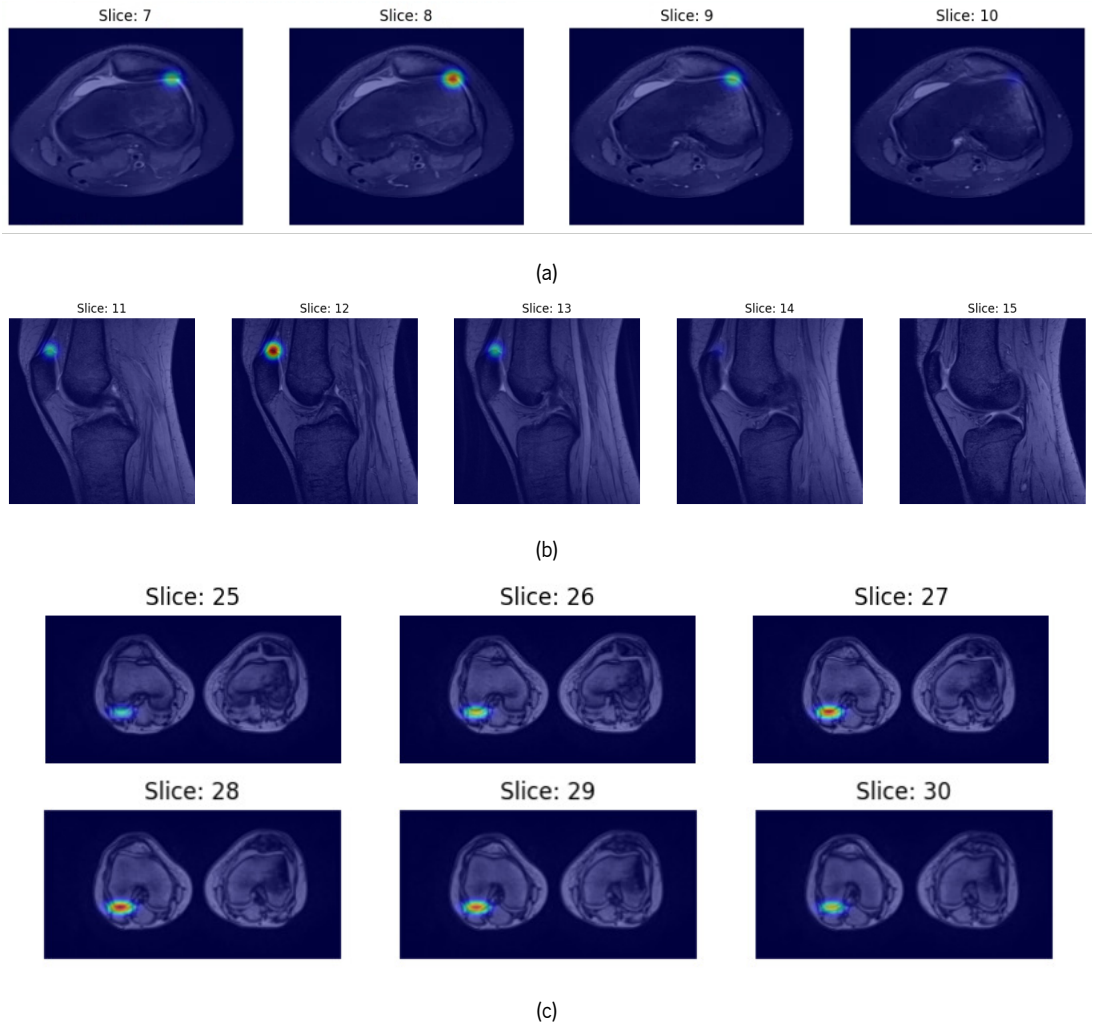


Figure 6.8: Ground truth mask for landmark 0 in the subsets: (a) Axial, (b) Sagittal, and (c) Dynamic.

The creation of the ground truth was done under a classification problem context, meaning that each voxel belonging to the heatmap had to be classified as being or not being the landmark it contained. Since the sequences had a considered number of noise voxels, the creation of a *background* class was implemented in Tan et al (2021) [107]. This *background* B_S, calculated using Equation 6.3, was composed of the difference between the background voxel's values set at 1 and the sum of all the heatmaps associated with the landmarks channels. The total number of channels is given by by $N_S + 1$, where N_S is the number of landmarks for each subset $S \in \{\text{AXIAL, SAGITTAL, DYNAMIC}\}$ with an additional channel for the *background*. Therefore it justifies a total of 8 channels for DATASET_AXIAL, 12 channels for DATASET_SAGITTAL and a total of 19 channels for DATASET_DYNAMIC previously mentioned.

$$B_S(x,y,z) = \begin{cases} \exp\left(-\left(\frac{(x-x_k)^2}{2\sigma_x^2} + \frac{(y-y_k)^2}{2\sigma_y^2} + \frac{(z-z_k)^2}{2\sigma_z^2}\right)\right), & k = 1,2,...,N_S \\ 1 - \sum_{i=1}^{N_S} B_S(x,y,z), & k = N_S + 1 \end{cases} \tag{6.3}$$

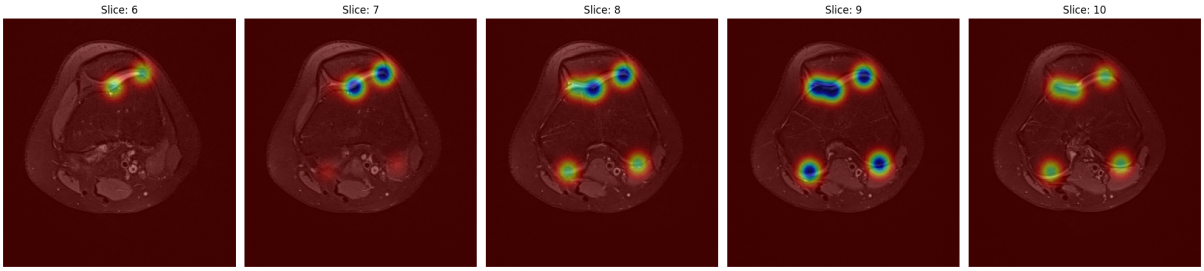


Figure 6.9: Ground truth mask for the background channel in the axial subset.

6.2.2 Resampling Volumes and Masks

Resampling was applied to both MRI image volumes and their respective segmentation masks to ensure dimensional consistency and facilitate processing and training. This preprocessing step was crucial for maintaining the integrity of spatial information during the fitting of DL models. To achieve this, the three dimensions of each subset had to be adjusted. These models, particularly those based on CNN, required input data that allowed for efficient downsampling, upsampling, and concatenation procedures, making dimensional consistency a key factor in the resampling process.

The resampling process involved the use of the SimpleITK library. Implementing a method capable of resampling each sequence for each subset was critical for the standardization of the process, obtaining a consistent and reliable source for fitting future models in the modeling stage. The *resample_sequence* function, using the SimpleITK library, was cleverly designed to adjust not only the volumetric data of the scans, but also the corresponding segmentation masks to a desired uniform size. This step was carefully designed to preserve the integrity of the labeled regions within the masks while maintaining their accuracy.

The process started by transforming the input MRI volume and masks into a SimpleITK-compatible format, enabling the application of advanced image processing techniques while preserving important spatial metadata. The core of the resampling logic focused on calculating the new spacing between voxels to achieve the target output size, while preserving the physical dimensions of the scanned anatomy.

One of the most critical decisions in the resampling process was the choice of interpolator, with the default function being linear interpolation, in this context Triliner interpolation (Figure 6.10). This choice is important because the interpolator determines how the intensity values of the new voxels in the resampled volume are calculated. Linear interpolation is a common choice in medical image processing, offering a compromise between computational speed and image quality.

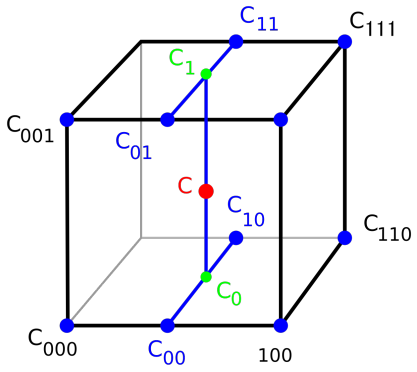


Figure 6.10: Depiction of Trilinear interpolation [108].

Given that the 3D U-Net architecture was the primary focus of experimentation in the pipeline, it was essential to ensure that the input dimensions were divisible by 32 (2^5). This requirement comes from the architecture's design, which includes five downsampling layers in the encoder and five upsampling layers in the decoder. Ensuring the input dimensions are multiples of 32 guarantees that the downsampling and upsampling processes result in the same resolution for both input and output, enabling voxel level loss calculation.

Consequently, the desired dimensions for width and height were set to a static value of 256 in all subsets, for both volume and corresponding channel's mask. The number of slices, however, varied between subsets: DATASET_AXIAL and DATASET_SAGITTAL were resampled to 32 slices, while DATASET_DYNAMIC, given its higher number of slices, was resampled to 64 slices.

Storage of Resampling Metadata

In parallel with the resampling process, it was necessary to store metadata from the original volume for each data subset. This data was crucial to the solution, as it allowed to revert the resampling of the test sequence/volumes, create for input into the model, back to the original size, after output from the model. This process was critical to obtaining valid metric results for comparison with literature results. For this purpose, the following metadata was stored in JSON format for each sequence:

- original_size and new_size;
- original_spacing and new_spacing;
- interpolation;
- pixel_spacing;
- slice_thickness.

6.2.3 File formats and I/O considerations

After the resampling stage, it was necessary to consider the final format in which all subsets would be stored. Initial efforts focused on the NIfTI (.nii) format, widely used for medical imaging due to its ability to store volumetric data and associated metadata. However, challenges with handling larger datasets, especially in terms of I/O performance, scalability and model fitting led to a transition to the TFRecords format. TFRecords offers several advantages such as efficient serialization, better performance, and scalability, making it more suitable for large-scale biomedical datasets.

TFRecords Database

The transition to TFRecords was implemented providing efficient data serialization. The format simplifies the way large datasets are managed, providing a number of valuable benefits to the medical imaging community. Using the TFRecords data format, TensorFlow's proprietary format for storing data of binary records, allows the training examples to be serialized and stored atomically on disk with fast write access and significantly reducing disk I/O overhead. The process involved encoding volumetric MRI data and associated masks into a compact binary format optimized for TensorFlow consumption allows the computational pipeline not to be susceptible by data-loading bottlenecks. However given the dimensions of the subsets it was trial-and-error process.

The pipeline for processing TFRecords involved several critical steps to ensure that the volumetric data and corresponding masks were properly formatted and normalized for DL models. The first step was parsing the TFRecords files. The serialized data, written in binary format, was read, converted into tensors, and reshaped to ensure the correct orientation for model consumption. The nibabel library played a key role in reshaping and transposing the arrays to ensure they were in the correct orientation. This parsing step was crucial because the way data was written to disk had to be mirrored in how it was read. The *parse_tfrecord* function meticulously decodes the serialized data into tensors that can be used by the DL models. This decoding is a crucial step in transforming the compactly stored data into a format that the models can interpret and learn from, thereby maintaining the integrity and usability of the data within the TensorFlow framework.

The TFRecords pipeline was designed to enable both cross-validation and non-cross-validation approaches. The *dataset_tfr_split* function facilitated the separation of training, validation, and test datasets. While the test dataset creation was not initially a primary consideration, it became necessary to separate the evaluation tasks from the modeling ones. This function laid the groundwork for an eventual approach to handling test datasets.

6.2.4 Data Normalization of Volume Voxel Intensity

As with natural images, normalization of the dataset is another critical component of the pipeline, akin to the preprocessing considerations in Deep Learning Toolkit for Medical Imaging's approach [109]. The process aim of normalization was to remove some variation in the data (e.g. differences in image contrast, etc.) that is known and so simplify the detection of subtle differences that would interest the solution in instead (e.g. the presence of a pathology).

Altough the choice between min-max and standard (z-score) normalization methods is offered, enabling the flexibility to tailor the preprocessing to the specific requirements of the model and the data distribution, the first was the only one used in the process. This normalization process, executed through functions, for instance *tf_min_max_normalize*, ensured that the input data to the model adhered to a consistent scale and that this data fed into the models was adjusted according to the specific distribution of the subset.

The normalization parameters were calculated using the *get_norm_params* function, which iterated over the dataset to determine the global minimum and maximum values of the voxel intensities. These values were essential for the min-max normalization, which scaled the voxel values to a range of [0, 1]. By transforming the data in this way, the network could more effectively learn the underlying patterns without being influenced by the original scale of the data.

To load and process the TFRecords, the *load_tfr_dataset* function was employed. This function read the TFRecords, applied the appropriate normalization method, and ensured the dataset conformed to the expected shapes. The normalization was applied dynamically during the data loading process, ensuring that the data fed into the model during training and evaluation was consistently normalized. The standardized input data facilitated the learning process of the DL models, improving their ability to generalize from the training data to new, unseen data.

Overall, for all the models fitted, the normalization process involved calculating the minimum and maximum intensity values from the training data. These values were used to normalize both the training and validation datasets, ensuring consistent input data across both phases. For the test dataset, the same minimum and maximum intensity values from the training data were used to normalize the data, ensuring unbiased and realistic evaluation. This process ensured that each model learned to normalize input data consistently, applying the same normalization parameters to both seen (training) and unseen (validation/testing) data.

6.3 Modeling Description

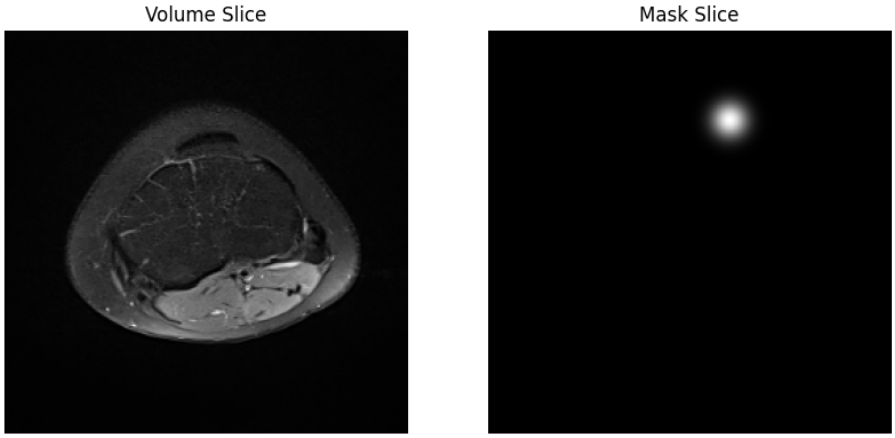
6.3.1 Data Augmentation

Data augmentation is a critical process for increasing the diversity of dataset and improving the generalization of the model. The developed augmentation functions were designed to be part of the TensorFlow graph, acting directly on the tensors and allowing efficient data manipulation independent of the data I/O format. In general, the implemented augmentations were applied in real time during training. This dynamic approach is efficient and does not increase the storage requirements of the training subset. The implemented augmentations were applied to the entire volume of images (3D scale) rather than to each slice/image individually (2D scale) to preserve the integrity of the volumetric information.

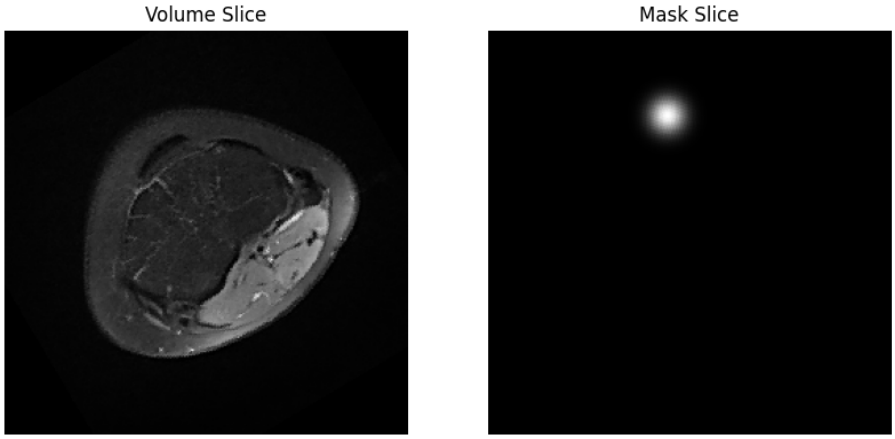
To incorporate variability and avoid deterministic transformations that the model could simply memorize, a stochastic component was introduced through *tf.random.uniform([])*. This function generates a random value, between 0 and 1 each time an image is processed, and if this value exceeds a predefined threshold, the augmentation process is applied. By using this component, the system decided when to use each augmentation function during model fitting.

Random Rotation

Volumes can be randomly rotated around the slice depth (z-axis) to mimic variations in patient positioning (Figure 6.11). The rotation angle is drawn from a uniform distribution in the range of approximately -0.7 to 0.7 radians (approximately -40 to 40 degrees), meaning every possible angle between -40 and 40 degrees has an equal chance of being selected. This range was chosen to simulate realistic variations in patient positioning that may occur during image acquisition. The TensorFlow Addon *tfa.image.rotate* operation was used to rotate each slice. This augmentation procedure was applied to the volume if a specific stochastic variable was greater than 0.3, with a 70% probability of occurrence. This augmentation was applied only apllied to Axial and Sagittal subsets.



(a)



(b)

Figure 6.11: Volume and mask: (a) before, and (b) after random rotation for a DATASET_AXIAL sequence.

Random Horizontal Flip

Images and their corresponding labels can be flipped horizontally (Figure 6.12), using the TensorFlow operation *tf.image.flip_left_right*, if *tf.random.uniform([])* > 0.3 (70% chance of occurrence). This simulates the variability between left and right legs. This augmentation is specific to the axial subset, since it was the only subset to which it was applied.

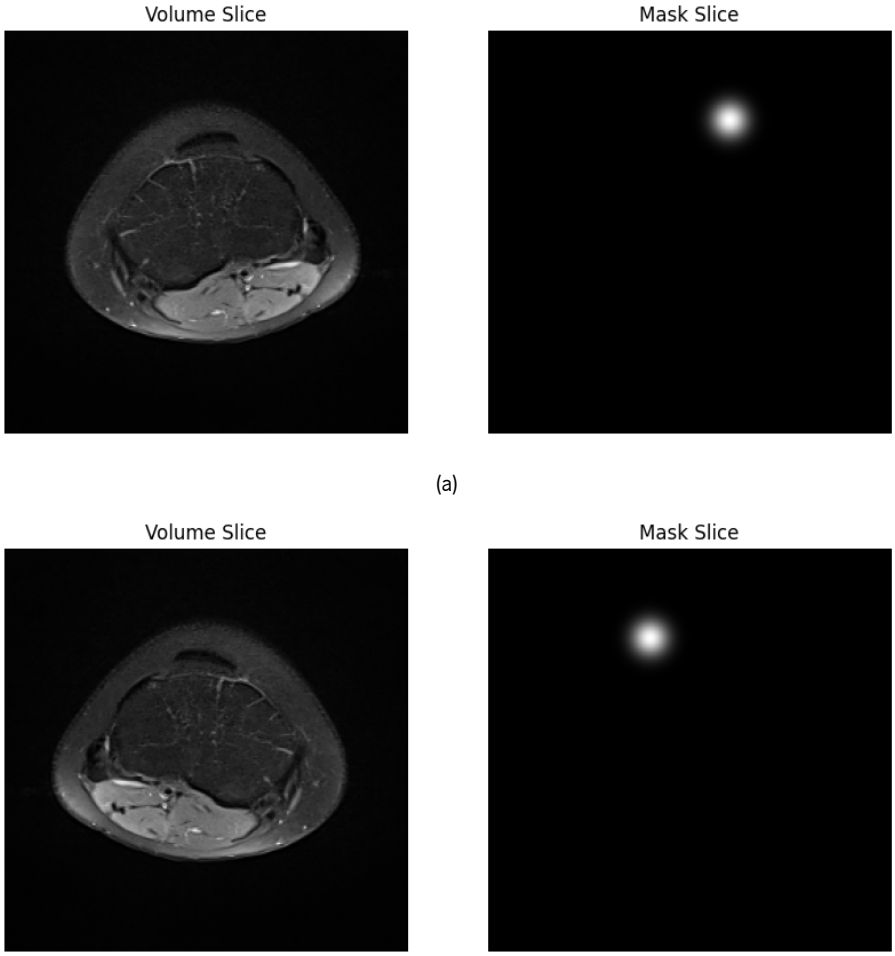
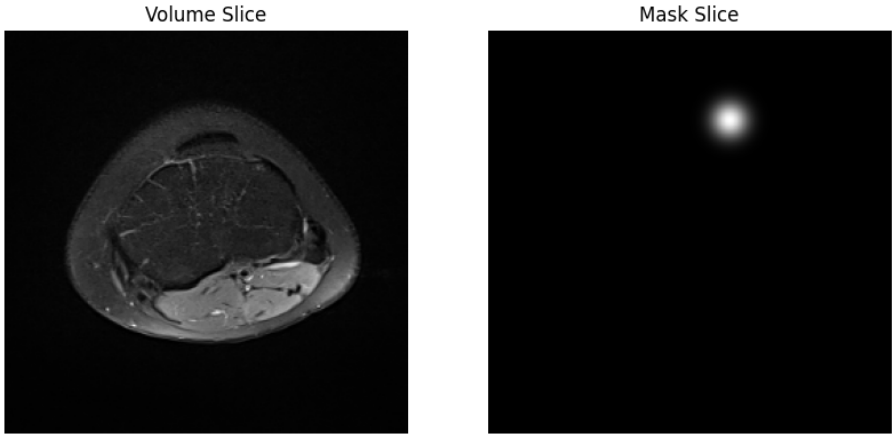


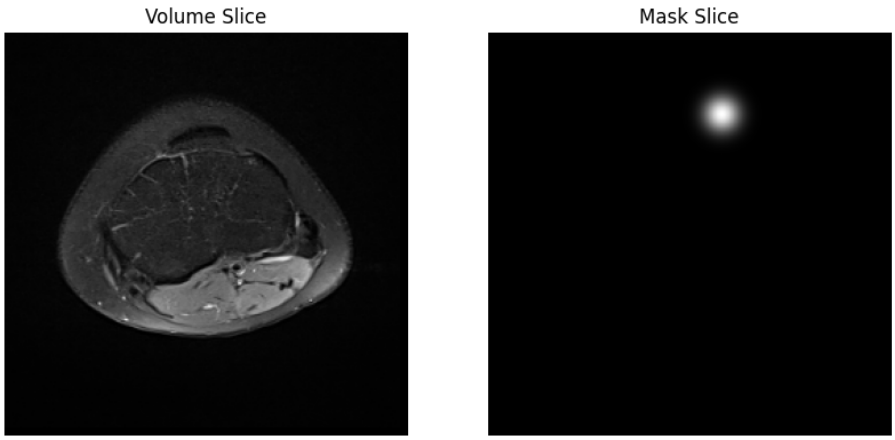
Figure 6.12: Volume and mask: (a) before, and (b) after random horizontal flip for a DATASET_AXIAL sequence.

Random Translation

Volumes are randomly shifted in the anteroposterior (front-back, in the x- direction) and left-right directions (y-direction) within a defined range, allowing the network to learn from different positional offsets in the data (Figure 6.13). The intensity of the translation was controlled by a *threshold* parameter, determining the maximum number of pixels each slice can be shifted in any direction. For instance, if the *threshold* is set to 10, each slice can be translated up to 10 pixels in the anteroposterior and left-right directions. This transformation is executed with a probability of 60% and involves the generation of random translations. The TensorFlow Addon *tfa.image.translate* operation was used to shift each slice, and this was performed consistently throughout the volume to preserve volumetric information. This augmentation was applied to all subsets.



(a)

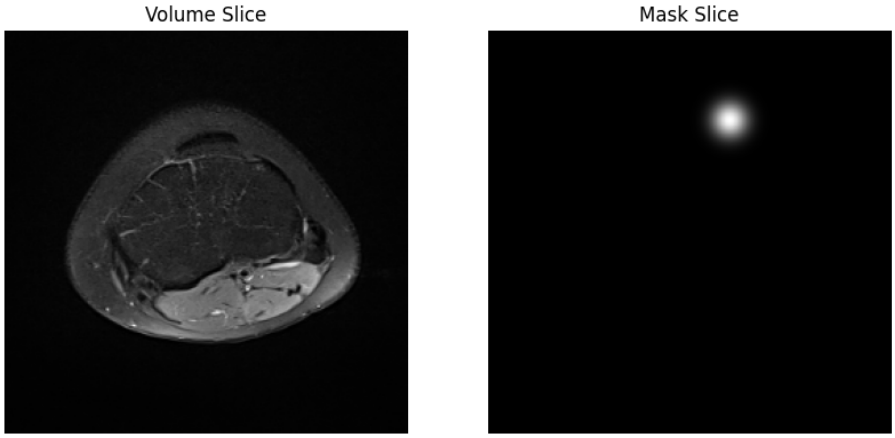


(b)

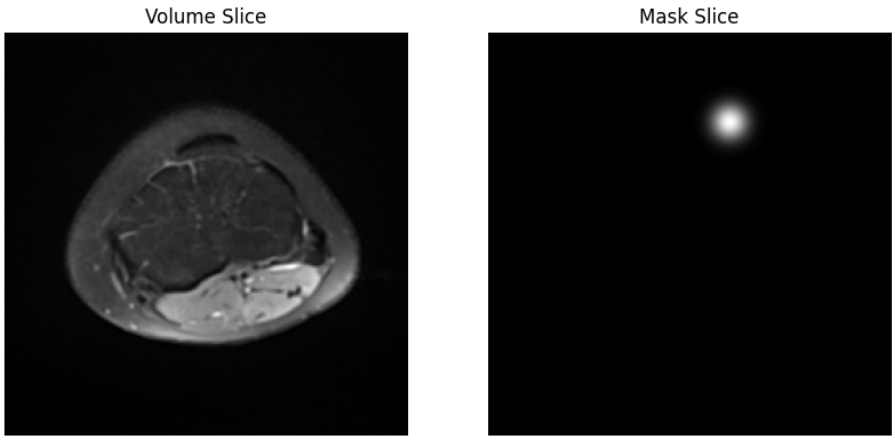
Figure 6.13: Volume and mask: (a) before, and (b) after random translation for a DATASET_AXIAL sequence.

Random Gaussian Blur

A Gaussian filter was also applied to, and only to the volumes to mimic the blurring effect that can occur due to patient movement or poor acquisition due to other factors (Figure 6.14). The intensity of the blur was controlled by a *sigma* variable, which was set to 2.0 by default. This enhancement was triggered when *tf.random.uniform([])* > 0.4 (60% chance of occurrence) and it was applied to each slice independently using *tfa.image.gaussian_filter2d*, which is useful for preserving the structural integrity of 3D volumes. This augmentation was applied to all subsets.



(a)

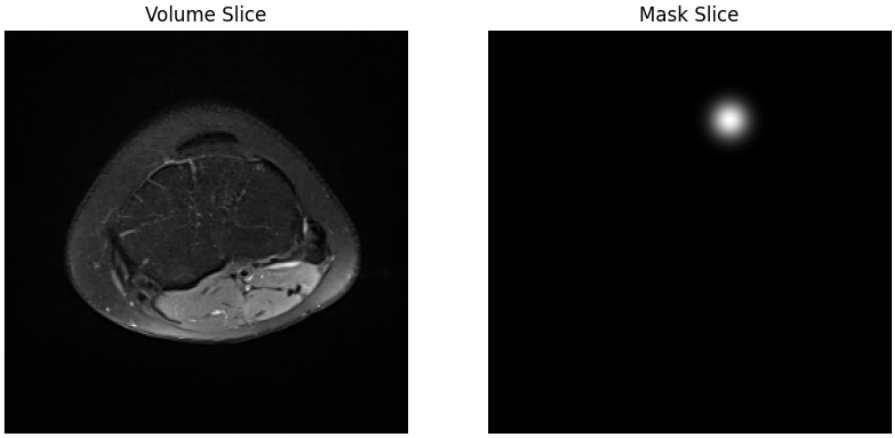


(b)

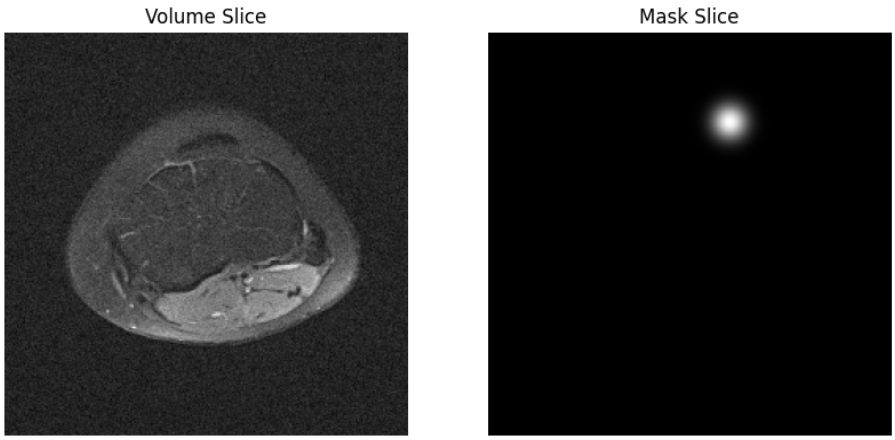
Figure 6.14: Volume and mask: (a) before, and (b) after random gaussian blur for a DATASET_AXIAL sequence.

Random Noise Injection

Gaussian noise was added to, and only the volumes to simulate the electronic "noise" present in MRI data (Figure 6.15). The intensity of the noise was determined by *mean* and *stddev* vairables. The standard deviation of the noise was randomly chosen within a defined range, with *mean* = 0.0 and *stddev* = 0.01, low-intensity Gaussian noise was added to the volume slices. As in the previous transformations if the random condition *tf.random.uniform([])* > 0.4 (60% probability of occurrence). This augmentation was applied to all subsets.



(a)



(b)

Figure 6.15: Volume and mask: (a) before, and (b) after random noise injection.

Using probability transformation maintains a balance between the original data and augmented data. This dynamic approach to data augmentation ensures a diverse training dataset, increasing the robustness of the resulting models against overfitting and allowing them to better generalize to new and unseen data. This probabilistic approach ensures that the model is exposed to both augmented and original images during training, enhancing its ability to generalize.

6.3.2 Deep Learning Models Architectures

The chosen DL models use deep CNN to process 3D MRI data. As mentioned earlier in Chapter 2, CNN are specifically designed to process data in a grid-like topology, such as images, and have proven to be highly effective in various detection tasks. By employing multiple convolutional layers, CNN excel at extracting and combining a variety of features to create detailed feature maps. These maps are pooled, activated, and combined to generate outputs, enabling CNN to recognize complex features and patterns in medical data.

These architectures were chosen based on their demonstrated success in the literature review conducted in Chapter 3. It concluded that while 3D data offers enriched spatial information crucial for nuanced landmark detection, it also requires sophisticated computational strategies and architectures to manage its inherent complexity. The architectures used in this study extend the conventional 2D U-Net model for a 3D version of it [106] emerging as a versatile and robust model. The network is

widely used for biomedical image segmentation, by introducing modifications to enhance its ability to capture complex patterns and variations in the data. Its architecture of encoder/decoder has been widely adopted for its robust segmentation capabilities and precise localization.

The DL models were implemented using TensorFlow, providing a comprehensive and flexible platform for building and training complex neural network architectures. TensorFlow's eager execution provides intuitive debugging and a natural control flow for developing, fitting and evaluating custom model architectures.

Each model reflects a progression in the complexity and capability of the architecture, from the construction of the simple 3D U-Net to the more sophisticated Residual Attention 3D U-Net. These models serve as the backbone of the segmentation approach, with each iteration building on the powerful TensorFlow framework for optimizing performance in the MRI volume segmentation task.

1. Simple 3D U-Net

The simple U-Net 3D model (Figure 6.16), constructed with the *unet3d* function, uses a standard U-Net architecture with classic convolutional operations for feature extraction and a series of downsampling and upsampling layers to capture context and locate segmentation areas. This model is the foundational approach, providing a baseline for convolutional network performance on volumetric data without advanced architectural enhancements.

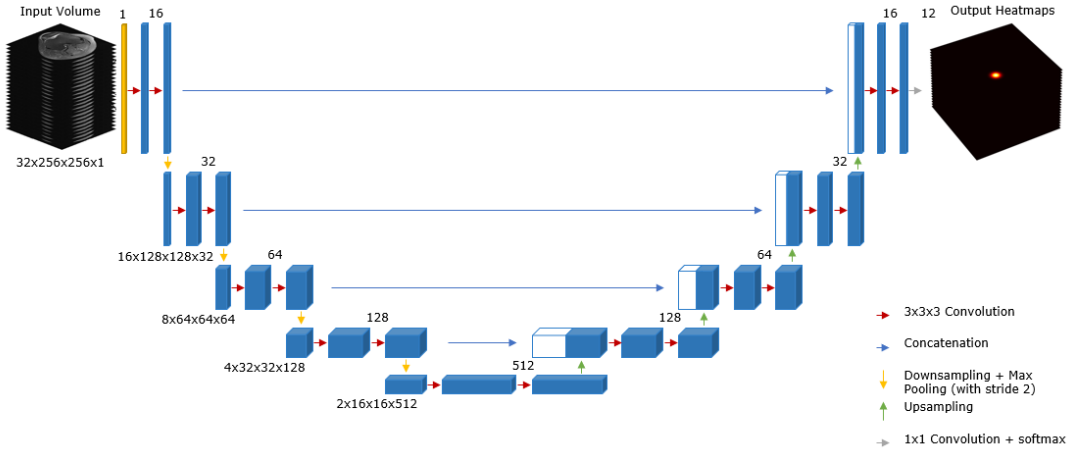


Figure 6.16: 3D Simple U-Net representation.

2. Residual 3D U-Net

The 3D Residual U-Net created by the *residual_unet3d* function introduces residual connections into its layers. These connections allow direct flow of information between layers, mitigating the degradation problem that can occur in deep networks and allowing deep architectures to be efficiently trained.

3. 3D U-Net with Attention Gates on Decoder Path

A U-Net with attention gates integrated into the decoder path was created using the *AttUnet3d* function. These attention gates allow the network to focus on salient features that are crucial for accurate segmentation, resulting in more precise localization of structures within MRI volumes.

4. Residual Attention 3D U-Net

The *resAtt_unet3d* function synthesizes the strengths of the previous model by combining residual blocks with attention gates, thus benefiting from the improved gradient flow of the residual connections and the selective focus of attention mechanisms. This architecture is designed to be robust, allowing for detailed segmentation while being deep and comprehensive in its learning capacity.

6.3.3 DL Architecture Methods and Layer Logic

In this subsection, we delve into the specific architectural methods and layer logic employed in our DL models. Understanding the implementation of each component is crucial for appreciating how these models effectively process and analyze 3D MRI data.

Normalization Two normalization techniques were considered, Batch Normalization (BatchNorm) and Group Normalization (GroupNorm), to stabilize learning and enable higher learning rates. BatchNorm remains the commonly used in DL operating by computing the mean and variance of each mini-batch to normalize the inputs. This technique helps mitigate internal covariate shift, allowing for faster convergence and the use of higher learning rates [110]. It also acts as a regularizer, reducing the need for other forms of regularization such as dropout. GroupNorm divides the channels of each layer into smaller groups and computes the mean and variance within each group for normalization [111]. This approach allows Group Normalization to perform consistently regardless of batch size, making it particularly useful in scenarios where batch sizes are inherently small or variable, such as in object detection or instance segmentation tasks.

Activation Function In this study, both Rectified Linear Unit (ReLU) and Leaky Rectified Linear Unit (Leaky ReLU) activation functions were used for intermediate layer processing. These activations played a crucial role by introducing non-linearity. During the fine-tuning process, ReLU consistently provided better results, leading to its selection as the primary activation function for the intermediate layers [112].

For the final classification layer, where the goal was to classify voxels in each channel, the Softmax activation function was employed. This choice was due to the need for a probabilistic output across multiple classes, ensuring that the sum of probabilities for all classes equals one. This activation function is particularly suitable for multi-class segmentation tasks, making it the optimal choice for this layer.

Additionally, the Sigmoid activation function was utilized in the attention mechanisms tested within the network. Sigmoid activation is well-suited for gating mechanisms as it outputs values between 0 and 1, effectively controlling the importance of different features by weighing them accordingly. This contributed to the chosen model's ability to focus on the most relevant features during the learning process.

Convolutional Layers The convolutional layers, implemented by the *convolution* function, acts as the main feature extractors in their models. They use standard and transposed convolutions to handle different aspects of network operation —standard convolutions for feature extraction and dimensionality reduction, and transposed convolutions for upsampling and dimensionality expansion. The kernel size for these convolutions is typically set to 3×3×3, which is a standard choice that balances computational efficiency and the ability to capture spatial features. The stride for downsampling operations is set to

2, effectively reducing the spatial dimensions by half, while upsampling operations also use a stride of 2 to double the spatial dimensions. Downsampling blocks, which belong to the encoder, and upsampling blocks, which belong to the decoder, are crucial for the U-Net architecture. They help to progressively reduce and then restore spatial resolution while extracting and refining features.

Two downsampling blocks, such as *downsample_block* and *downsample_Residual_block* were constructed to to reduce the spatial dimensions and increase the number of feature channels. The last block differs the first by including a direct shortcut path that bypasses the main layers, these blocks allow gradients to flow more freely through the network, mitigating the vanishing gradient problem common in deep architectures. This is achieved by adding a 1x1x1 convolution with a stride of 2 to match the dimensions of the main path, ensuring that the shortcut connection is compatible with the output of the convolutional layers within the block.

Four upsampling blocks, including *upsample_block*, *upsample_Residual_block*, *upsample_attention_block*, and *upsample_ResidualAttention_block* employ transposed convolutions to restore the spatial dimensions downsampled on the encoder part. As in *downsample_Residual_block*, the same process of direct shortcut is applied to *upsample_Residual_block* with the residual connections. Moving to attention mechanisms, implemented in the *attention_gate* and *upsample_attention_block* functions, implemented gates allowing the model to selectively focus on the most relevant features of the data. By weighing the feature maps from previous layers, the network can prioritize regions of interest over less relevant areas, leading to better performance in segmentation tasks where precise spatial details are crucial. The integration of these attention mechanisms within the residual blocks, as seen in *upsample_ResidualAttention_block*, further enhances the network's ability to focus on relevant features and maintain stable gradient flow.

Furthermore He initialization is used for the kernel initializer in the convolutional layers. This method is particularly suitable for layers with ReLU activation functions, as it helps to maintain a healthy variance of the input signals throughout the network. Although regularization techniques like L2 regularization were considered, the code indicates that these were not used. Instead, the models relied on the inherent robustness of the chosen architectures and the regularization effect of the normalization techniques employed.

Loss Functions and Optimizers For the segmentation tasks in this project, two loss functions were considered: Dice loss and the composite loss function Dice+CE. Dice loss focuses on maximizing the overlap between the predicted and true segmentation masks, which is particularly effective in handling class imbalance in segmentation tasks. The composite loss function Dice+CE, which combines Dice loss and Categorical Cross-Entropy (CE) loss, was chosen as it achieved better results. This composite loss leverages the specificity of Cross-Entropy loss for accurate voxel classification and the spatial alignment focus of Dice loss. This combination ensures a balanced approach that addresses both class imbalance and the need for precise spatial segmentation. Finally the Adam optimizer complements this by efficiently handling gradient variability and adjusting learning rates to promote the development of robust segmentation models.

Callbacks During model training, several callbacks were utilized to enhance the training process. These callbacks included several important topics: Model Checkpoint in order to save the model at specific intervals or when the validation score improved, ensuring that the best model could be retained even if the training was interrupted; *EarlyStopping* to prevent overfitting by monitoring the validation loss and stopping the training early if the loss did not improve for a defined number of epochs; a learning rate scheduler, *ReduceLROnPlateau* important to adjust the learning rate dynamically during training, capable of lowering it when progress plateaued to fine-tune model weights; *TerminateOnNaN* callback was employed to terminate the

training process if a NaN (Not a Number) or infinite loss was encountered, preventing the continuation of training with invalid loss values; the *LearningCurveSaver* callback was used to save the learning curve data, including training and validation loss, to a JSON file at the end of training. This allowed for detailed assessment of the model's performance over time and identification of potential areas for improvement.

In the modeling notebooks for the AXIAL, SAGITTAL, and DYNAMIC subsets, the callback setups demonstrate both uniformity and specific adaptations to meet the unique needs of each subset. Common and static to all three are the *TerminateOnNaN*, *ModelCheckpoint*, *TerminateOnNaN* and *LearningCurveSaver* callbacks, which provide a robust framework for monitoring training progress, safeguarding against non-convergent training iterations, and ensuring that the best model states are preserved. Differences arise primarily in the patience settings for the *EarlyStopping* and *ReduceLROnPlateau* callbacks. Given that the models tune for each data subset could be different, these specifications should maintain consistency in execution while adapting to the distinctive characteristics of each dataset.

6.4 Conclusion

The chapter not only outlines the technical processes involved in preparing and processing data, but also emphasizes the importance of a methodical approach to data in medical imaging. The preprocessing section, covering ground truth creation, resampling and normalization, was pivotal in ensuring the data was consistently and effectively prepared. The exploration of each step highlights how structured data management and what innovative modeling techniques can be valid study options for medical diagnostics. The modeling section demonstrated the implementation of advanced CNN designed to handle the complexity of 3D MRI data, using a combination of data augmentation techniques and sophisticated architectural enhancements such as attention gates and residual connections to improve the model performance in segmentation prior to localization. The use of dynamic data augmentation was instrumental to enriching the diversity of the dataset and improving model generalization. Overall, the chapter demonstrated the potential of these advanced methods applied to a more specific case of medical image images.

Chapter 7

Results and Discussion

This chapter presents a comprehensive analysis of the results obtained from the various models and subsets used in the study. The evaluation framework was designed to measure model generalization and landmark detection performance on the different axial, sagittal, and dynamic data subsets. The chapter discusses the division of training and test subsets, the training strategies employed, the metrics used to assess model performance, the results obtained, and the benchmarking of these results against those of other studies. It also explores the impact of different normalization techniques and architectural variations on prediction accuracy. The results are analyzed in terms of millimeter-based errors to provide a detailed understanding of the models' effectiveness in clinical contexts. Finally a Benchmarking section is provided for comparing the projects results with state of the art results.

7.1 Cross Validation Approach

For each subset, the data was divided into training and testing sets, with a primary focus on creating a representative test dataset. The subjects were divided into 80% for training and 20% for testing. Of the 80% allocated for training, the same process was applied, resulting in 80% for training and 20% for validation. Each subject only entered once, either in training, testing, or validation, which prevented the training and evaluation from being biased. After constructing the test sets, the remaining data was used for training. A 5-fold cross-validation was employed to further split the training data into training and validation sets. This method involved dividing the data into five partitions, training a selected model on four partitions, and validating it on the remaining one. This process was repeated five times, each time with a different partition used for validation. By averaging the validation scores across all five folds, a more reliable evaluation of the model's performance was obtained [113]. This approach minimized the variance in validation scores and ensured a robust assessment of the model. The full cross-validation process was only applied to the Axial and Sagittal subsets. The use of 5-fold cross-validation proved helpful in the fine-tuning process. Given that the seed value under each subset's notebook was the same, the division of training/validation data was repetitive and allowed the use of one and specific fold for fine-tuning.

In the Axial subset, comprising a total of 234 sequences, 48 sequences (20.51%) were allocated to the test dataset. From the remaining 186 sequences, 149 (63.68%) were assigned to the training dataset, and the remaining 37 (15.81%) were used for validation. For the Sagittal subset, which included 429 sequences, the test dataset comprised 84 sequences (19.58%). From the remaining 345 sequences, the training dataset included 276 (64.34%), and 69 sequences (16.08%) were used for validation. Lastly, the Dynamic subset, with a total of 330 sequences, was divided such that 68 (20.61%) of the sequences were used for the test dataset. Out of the 262 sequences available for training, 209 (63.33%) were assigned to the training dataset, and 53 sequences (16.06%) were utilized for validation.

7.2 Test Dataset Composition

The test dataset is a critical component of the evaluation framework, serving as the basis for measuring the model's generalization and diagnostic performance on unseen data. This stage should be kept separate from training and validation datasets to prevent data leakage and ensure that the model's performance metrics are reliable and generalizable to new data. A different test dataset was created for each subset, attempting to balance and maintain the actual proportions of healthy and pathological cases. In addition, the proportions of males and females, and left and right leg sequences were taken into account. In this way, it was possible to represent a balanced cross section of the patient population regarding the specifications of each subset, including a diverse range of cases to ensure a comprehensive test. A total of 18 subjects were selected for each subset of axial, sagittal and dynamic test data. Each subject is classified as either healthy or pathological.

7.2.1 Axial Test Subset

Table 7.1 shows the subjects that are part of the DATASET_AXIAL test subset.

Table 7.1: Composition of the Axial test dataset

Diagnosis	Patient ID	Gender	Left Sequences	Right Sequences
Healthy	5	M	1	2
	20	F	1	1
	23	F	1	1
	35	M	2	0
	47	M	2	2
	56	F	2	0
	66	M	2	2
	72	F	0	2
	77	M	0	1
	82	F	0	1
Pathological	12	F	2	0
	13	F	3	2
	39	F	2	1
	44	M	2	2
	46	F	2	2
	51	F	2	0
	64	M	3	0
	70	M	0	2
Total	18	-	27	21
			48	

M: Male, F: Female

The demographic distribution of the axial test data presents a balanced gender representation with a slight female predom-

inance. The healthy group consists of 5 males and 5 females, while the pathology group consists of 3 males and 5 females. The laterality of the sequences indicates the distribution of left and right knee images in the dataset. While there are 11 left and 12 right sequences in the healthy group, there are 16 left and 9 right sequences in the pathological group. A total of 48 axial sequences were used to test the model.

7.2.2 Sagittal Test Subset

Table 7.2 shows the subjects that make up the DATASET_SAGITTAL test subset.

Table 7.2: Composition of the Sagittal test dataset

Diagnosis	Patient ID	Gender	Left Sequences	Right Sequences
Healthy	5	M	3	3
	20	F	3	3
	23	F	3	3
	34	M	3	0
	45	M	3	3
	54	F	3	0
	64	M	3	3
	70	F	0	3
	75	M	0	3
	81	F	0	3
Pathological	12	F	3	0
	13	F	2	3
	38	F	3	3
	42	M	3	4
	44	F	3	3
	49	F	3	0
	62	M	6	0
	68	M	0	3
Total	18	-	44	40
			84	

M: Male, F: Female

The sagittal test dataset data also has a balanced distribution in terms of gender and left and right knee images, ensuring robustness in different anatomies. The healthy group consists of 5 males and 5 females, while the pathology group consists of 5 females and 3 males. Regarding the distribution of images of the left and right knee, in the healthy group there are 21 and 24 sequences, respectively. The pathology group shows a slight variation with 23 left and 16 right sequences. A total of 84 sagittal sequences were then used to test the model.

7.2.3 Dynamic Test Subset

Table 7.3 lists the subjects that comprise the DATASET_DYNAMIC test set. Due to the nature of the dynamic sequences, which include both left and right knees in a single volume, the number of left and right sequences is always equal.

Table 7.3: Composition of the Dynamic test dataset

Diagnosis	Patient ID	Gender (M = Male, F = Female)	Sequences
Healthy	5	M	4
	20	F	4
	23	F	4
	33	M	4
	43	M	4
	51	F	4
	61	M	4
	67	F	4
	71	M	4
	76	F	4
Pathological	12	F	4
	13	F	4
	37	F	4
	41	M	1
	42	F	3
	47	F	4
	59	M	4
	65	M	4
Total	18	-	68

M: Male, F: Female

The dynamic test dataset is characterized by its focus on capturing the dynamics of knee movements. The datasets presents a gender balanced representation with a total of 18 patients divided into 8 male and 10 female subjects. Due to laterality terms, the healthy group with 40 sequences has 40 left knees and 40 right knees, and the pathological group with 28 sequences has 28 left knees and 28 right knees. A total of 68 dynamic sequences were then used to test the model, with 68 left and 68 right knees.

7.3 Evaluation Process

After training, it is important to evaluate the performance of a model. As with the training of the models, the Axial, Sagittal, and Dynamic subsets were evaluated independently.

The process of evaluatig all subsets required the same TFRecords data functionalities used to build the training and validation datasets. The data handling process began by iterating through the test subject's paths, followed by the creation of

a *tf.data.Dataset* object. This was used to make predictions on the sequence volumes, which were then processed to obtain the Center of Mass (CoM) of the predicted heatmaps. These values were then compared with the CoM from the ground truth masks. With these comparative data points, it was possible to perform evaluations on all the specific channels/landmarks pertinent to each subset.

In the context of regressing the predicted heatmaps to a 3D coordinate system of landmarks, a specific methodology was studied and implemented to obtain the CoM. The *find_CoM* function was implemented, which identifies the CoM based on the peak values of the heatmap. This method locates the maximum intensity value within the specified channel and identifies the voxel positions corresponding to this peak value. If a single maximum value is found, its coordinates are returned directly. If multiple voxels share the maximum value, the mean position of these voxels is calculated and the resulting coordinates are rounded to the nearest integers for precise voxel location. This peak-based approach ensures that the most significant voxel positions are considered for accurate landmark detection. The pipeline for this process is as follows:

- Heatmap Channel Extraction: Extracts the specific channel from the 3D heatmap.
- Maximum Value Identification: Identifies the maximum intensity value within the extracted channel.
- Peak Value Location: Generates a coordinate grid for each voxel.
- Mean Position Calculation: Calculates the mean position of these voxel coordinates, if multiple peak values exist.
- Rounding and Type Conversion: Rounds the weighted coordinates to the nearest integers.

7.3.1 Evaluation Metrics

There were several valid options to evaluate this study, as as the review in Chapter 3 shows. Across all the studies in the review, it is possible to identify a set of regression and classification metrics. However, classification metrics were not used because of the clinical aspect required to effectively assess the results. For example, the use of SDR would require a preliminary study to define the acceptable range of error within which a landmark is considered to be accurately detected.

Mean Absolute Error

Among the reviewed metrics in Chapter 3, MAE stands as one of the most widely used. MAE was selected over Mean Error to avoid the potential for error cancellation, where positive and negative errors could offset each other, leading to an inaccurate representation of model performance. By considering only the absolute magnitude of errors, MAE provides a more accurate and consistent measure of prediction accuracy, equation 7.1,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7.1}$$

where y_i represents the ground truth values and $\hat{y}_i$ represents the predicted values, with n being the number of samples in the test dataset. This metric is commonly used under this type of DL problems and presents a robust way to demonstrate the results of the several subsets and models.

Loss

Loss, which played a critical role during the training phase, guiding the optimization of model parameters, was another metric chosen. By continuously monitoring and minimizing the loss, it was ensured that the models converged effectively, increasing their predictive accuracy and contributing to the development of a robust and generalized model. Three different types of losses were tried when fine-tuning the models, tested fot the subset DATASET_AXIAL. The process ended with the discovery of the optimal loss function used for the entire training approach for the three subsets.

Categorical Cross-Entropy (CE) Loss is commonly used for multi-class classification problems. The CE equation, 7.2, measures the performance of a classification model whose output is a probability value between 0 and 1,

$$\text{CE loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{\text{true},i,c}\log(y_{\text{pred},i,c}) \tag{7.2}$$

where N is the number of samples, C is the number of classes, $y_{\text{true},i,c}$ represents the ground truth label for class/landmark c of sample i, and $y_{\text{pred},i,c}$ represents the predicted probability for class c of sample i.

Dice Loss is used to assess the similarity between two sets. This type of loss is often used in segmentation tasks, which makes the 3D approach useful. The Dice equation, 7.3, measures the overlap between the predicted masks and the ground truth masks,

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C}(y_{\text{true},i,c}\cdot y_{\text{pred},i,c})}{\sum_{i=1}^{N}\sum_{c=1}^{C} y_{\text{true},i,c} + \sum_{i=1}^{N}\sum_{c=1}^{C} y_{\text{pred},i,c}} \tag{7.3}$$

where $y_{\text{true},i,c}$ represents the ground truth gaussian mask for for class/landmark c of sample i, and $y_{\text{pred},i,c}$ represents the predicted gaussian mask for c of sample i.

Constructing a combination of both Dice Loss and CE Loss, Dice+CE, leveraged the strengths of both metrics. The Dice+CE loss, equation 7.4, improves the overlap between predicted and ground truth masks while ensuring probabilistic predictions.

$$\text{Dice + CE Loss} = \text{CE} + \left(1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C}(y_{\text{true},i,c}\cdot y_{\text{pred},i,c})}{\sum_{i=1}^{N}\sum_{c=1}^{C} y_{\text{true},i,c} + \sum_{i=1}^{N}\sum_{c=1}^{C} y_{\text{pred},i,c}}\right) \tag{7.4}$$

7.4 DATASET_AXIAL

7.4.1 Training Approach

Most model tuning was conducted using the Axial subset, where all four architectures from Section 6.3 were created. Initial fitting began on Computer 1, using a unique fold from the 5-fold cross-validation approach mentioned earlier, for training. The model was tuned using a fixed seed value of 42 to ensure reproducibility. This approach, although not ideal, reduced training time and simplified testing. Consequently, the following loss and MAE results will only be for one fold.

The process started with a simple 3D U-Net. GPU memory caching was considered, but proved ineffective due to frequent Out-Of-Memory (OOM) errors. It proved more feasible to keep volumes in 4D format (depth, height, width and the 12 corresponding masks) and training without using cache memory. Another important aspect in solving these errors was the assessment of different batch sizes. The task started with a batch size of 16, which led to OOM errors on the GPU. Reducing the batch size to 4 mitigated some issues, but did not completely solve the problem. A batch size of 2 proved to be stable and was subsequently used consistently across all subsets. The training of DATASET_AXIAL started with a small number of epochs, ranging from 10 to 30, in a process of trial and error while encountering OOM errors. Later in the process, Computer 2 was used due to its availability.

During this stage, the first training loss experimented was the CE loss, due to the multi-class classification nature of the problem. This type of loss made the models converge quickly. However, the initial results were unsatisfactory. When visualising the resultant masks heatmaps, there was no clear indication of 'learning'. As the chosen architectures were mainly used for the segmentation of anatomical areas, and the ground truth consisted of gaussian heatmaps, Dice loss was subsequently tested. This change resulted in significant improvements, with the predicted heatmaps showing meaningful signs of learning. Encouraged by the positive results of the Dice loss, a more complex loss function, Dice+CE loss, was tested and ultimately used for the entire pipeline and all subsets. This loss provided a balanced approach to the classification and segmentation requirements of the problem, resulting in better convergence and prediction accuracy.

Once the correct loss function was found, the hyperparameters such as activation functions, convolution kernel sizes, learning rates, output feature maps for each layer, and layer normalization methods were tuned. Of these parameters, the convolution kernel sizes were considered the least relevant, and it was decided to leave them unchanged. Different activation functions were tried, specifically in the intermediate layers, while the final output activation function was always Softmax due to its suitability for multi-class classification. Both Leaky ReLU and ReLU were tested as activation functions in the intermediate layers. ReLU ultimately achieved better results, producing more accurate and precise predicted heatmaps, and was adopted for the entire pipeline in all subsets. The learning rate was set to 1×10^{-4}, after preliminary experiments with a learning rate of 5×10^{-5}, which was ultimately dropped due to slower convergence and suboptimal results. The learning rate of 1×10^{-4} balanced fast convergence and stable training dynamics.

The number of output feature maps is also important, particularly in architectures like U-Net. The number of layers and filters has been carefully tuned, as either too many or too few can lead to sub-optimal performance. Too many layers or filters could lead to overfitting, where the model becomes too complex and memorizes the training data instead of generalizing to unseen data. From the base architecture in [112] the default U-Net had about 35,519,756 trainable parameters, which was something unfeasible for the range of GPU and the nature of the input to the network. It was therefore necessary to reduce the number of layers and investigate the number of filters for each layer through an iterative refinement discussed in the loss section.

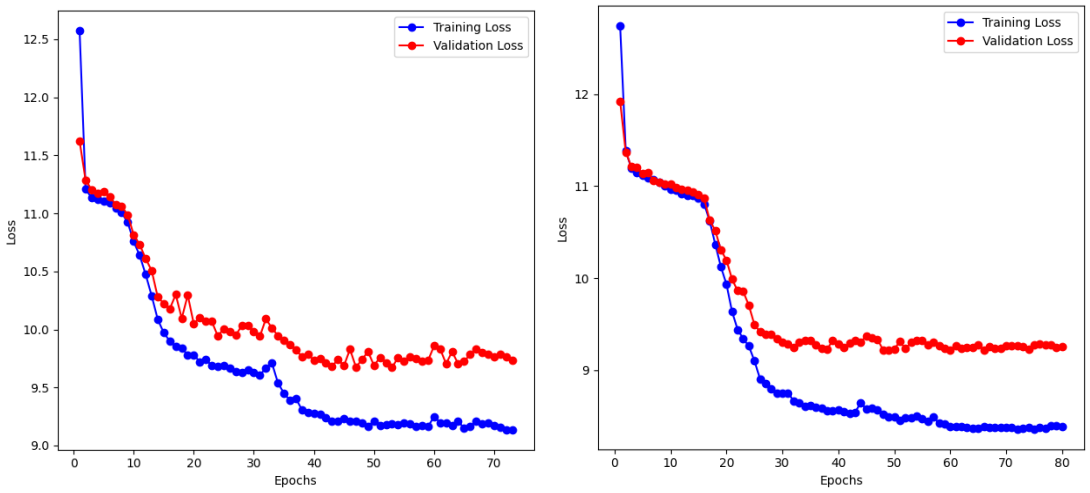
The final hyperparameter studied was the normalization technique. BatchNorm relied on large batch sizes to obtain statistically significant means and standard deviations, which can be problematic for 3D segmentation due to hardware limitations. Therefore GroupNorm was considered to the Axial dataset. Siince the output includes 12 channels/landmarks (including the background), GroupNorm was applied along the axes of the 3D vector and grouping its channels into sets. Specifically, introducing a hyperparameter G that determines the number of groups [114]. 4 groups were chosen to balance computational efficiency and maintain independent statistics within each group.

To further refine the training process, the learning rate scheduler was managed. Initial *ReduceLROnPlateau* settings aimed

to reduce the learning rate by a factor of 0.5 if the validation loss did not improve over a specific number of consecutive epochs, with a minimum threshold of 1×10^{-6}. For *EarlyStopping* the patience value was adjusted, starting at 10 and increasing to allow more epochs for slower convergence. The training went through several attempts to get the best results, combining the different callback parameters. Overall, the results presented were a product of the factor in *ReduceLROnPlateau* and patient parameters values for both callbacks. At this point the designated number of training epochs was between 90 and 100.

7.4.2 Loss

The first model to report was the simple 3D U-Net with [32,64,128,256] skip connections and a 320 bridge for the encoder, named *unet3d_ubuntu*, fitted on Computer 2 (Figure 7.1a). This architecture had approximately 18,928,076 trainable parameters. Next, a reduced version with [16,32,64,128] skip connections and a 512 bridge for the encoder (Figure 7.1b), named *unet3d*, was tested. These feature channels remained the default for other architectures. The validation loss decreased from *unet3d_ubuntu* to *unet3d*, with *unet3d*'s loss decreasing more steadily converging after more epochs.

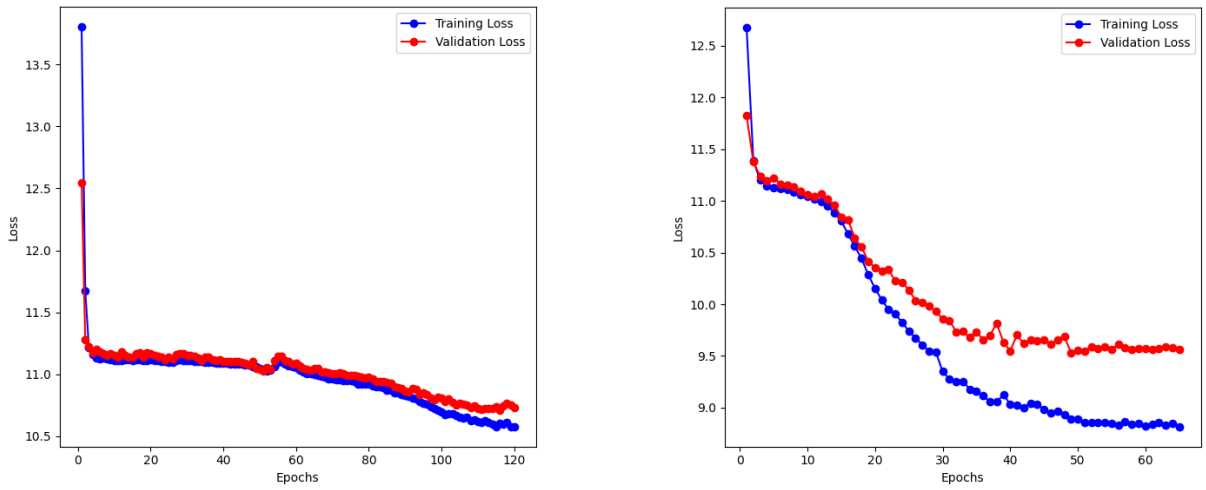


(a) (b)

Figure 7.1: Loss learning curves for Simple 3D U-Net using BatchNorm: (a) *unet3d_ubuntu*, (b) *unet3d*.

It is expected that some gap exists between the training and validation loss learning curves, with a smaller gap indicating better generalization. Both models exhibit some degree of overfitting, as indicated by the gap between training and validation losses. The *unet3d* model, however, shows a smaller gap, suggesting it generalizes better to unseen data. Comparing both *unet3d* models shows that the reduced feature channels perform better, suggesting that reducing feature channels may be beneficial for this task.

Residual connections and attention mechanisms were incorporated into the simple U-Net 3D architecture to evaluate their impact. Both of these architectures used [16,32,64,128] skip connections and a 512 bridge. The number of epochs for each training session at this stage were 120 and 100, for *residualunet3d* and *attunet3d* respectively . Through the analysis of the loss curve of the Residual U-Net, named after *residualunet3d* (Figure 7.2a) it is observed that the training loss starts at a higher value and rapidly decreases within the first few epochs, stabilizing as training progresses. The validation loss follows a similar trend, indicating good generalization. In the Attention based U-Net, named *attunet3d* (Figure 7.2b) the dropout regularization technique was used with a factor of 0.2.



(a) (b)

Figure 7.2: Loss learning curves for Simple 3D U-Net with residual connections and Simple 3D U-Net with attention mechanisms using BatchNorm: (a) *residualunet3d*, (b) *attunet3d*.

Further analysis shows that *attunet3d* converges faster than *residualunet3d*, reaching lower loss values in fewer epochs. On the other hand the *residualunet3d* actually uses the entire 120 epochs value detailed on the fitting call, with no need for early stopping. This suggests that the attention mechanism effectively helps the model focus on relevant features, speeding up the learning process. However in terms of generalization, *residualunet3d* demonstrates better generalization as indicated by the smaller gap between training and validation loss, despite achieving slightly higher final loss values. It is assessed that both residual connections and attention mechanisms improve the learning process of the U-Net 3D architecture in specific details, such as generalization. However, at the moment the simpler *unet3d* obtained better validation loss values, achieving convergence close to epoch 80.

Finally, the Residual Attention U-Net is reached, named *resattunet3d* (Figure 7.3). In this architecture both residual connections and attention mechanisms were combined, into the simpler U-Net. As the previous one, this architecture had [16,32,64,128] out channels for the skip connections and a 512 filter bridge. The *resattunet3d* model exhibited stable loss curves with less fluctuation in validation loss compared to all other architectures. It was possible to identify that 120 epochs was too large a value.

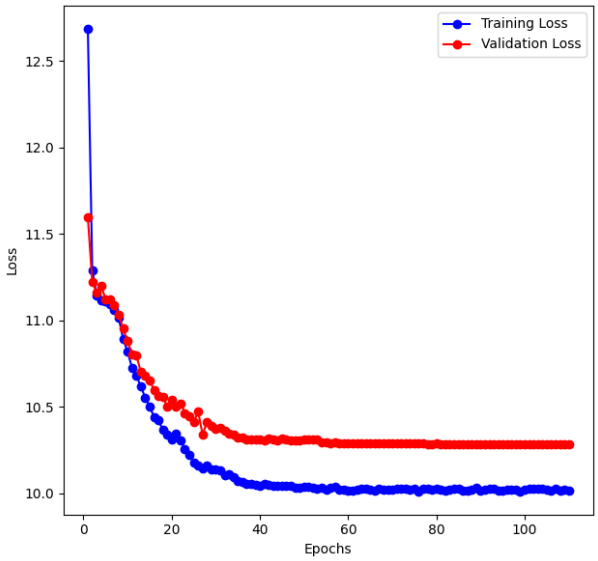


Figure 7.3: Loss learning curves for Simple 3D U-Net with residual connections and attention mechanism, *resattunet3d* using BatchNorm.

At the conclusion of the Axial loss stage, the learning loss histories for all architectures (*unet3d*, *residualunet3d*, *attunet3d*, and *resattunet3d*) were displayed. Although some details varied, the differences in final validation loss values were crucial for understanding the model tuning, learning process and the impact of different technologies on the Axial subset. Therefore the MAE evaluation was applied to all the architectures. When comparing GroupNorm to BatchNorm, GroupNorm did not consistently yield better results. While GroupNorm's performance was not significantly better than BatchNorm, it was also not considerably worse.

7.4.3 MAE

While assessing loss learning rate proved useful for tuning the models, the evaluation procedure was complemented with the use of MAE. The average MAE values for each landmark, for BatchNorm are presented in Figure 7.4. An average MAE value for each axial model across all landmarks is displayed (Table 7.4). All measurements are in millimetres (mm).

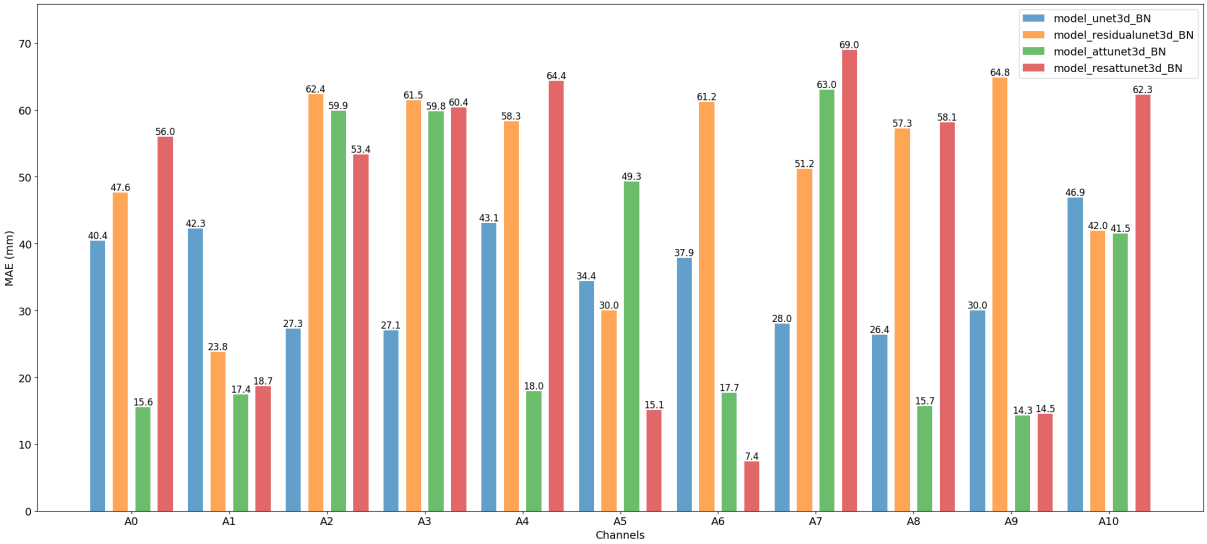
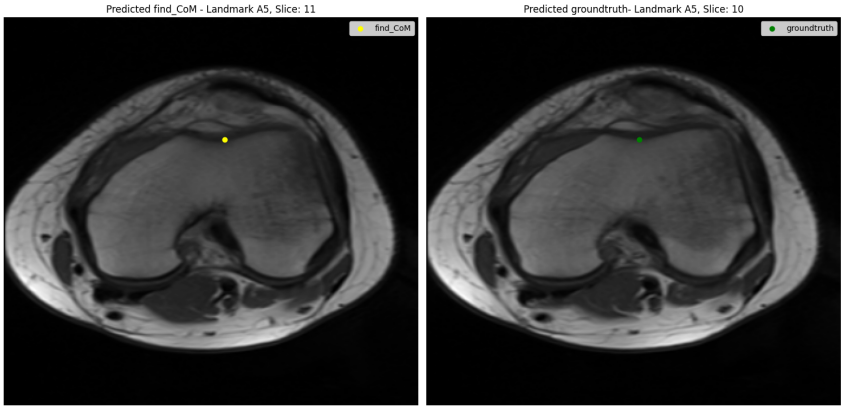


Figure 7.4: MAE for each of the architectures, for each of the landmarks in the Axial subset, in mm, using BatchNorm.

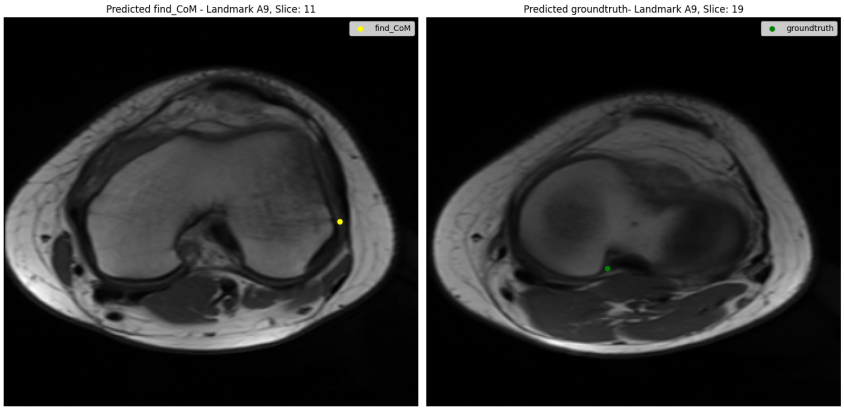
Table 7.4: Average MAE (mm) value for each axial model (across all landmarks) using BatchNorm

Architecture	**MAE** Mean±Std (mm)
model_unet3d_BN	30.99 ± 20.34
model_residualunet3d_BN	50.92 ± 13.15
model_attunet3d_BN	33.84 ± 19.84
model_resattunet3d_BN	43.57 ± 22.88

The assessment reveals a very unbalanced set of results for each landmark, which are quite unsatisfactory. The loss curves and the fact that the validation loss did not decrease as expected indicate that the models applied to DATASET_AXIAL need significant improvement. This includes not only further tuning but also revisiting previous steps. Therefore the majority of these unsatisfactory results can potentially be explained by the preprocessing stages. A more thorough investigation of the resampling process and ground truth construction is necessary. The descriptive analysis in Section 6.1 shows that DATASET_AXIAL had an average depth of 30 slices, ranging from 22 to 40, with high variance in slice spacing. Therefore, 32 slices were chosen to match the mean depth and ensure compatibility with the architectures. The section also notes a high variance in slice spacing, which is an important factor to consider. A finer pixel spacing combined with larger slice spacing might improve in-plane resolution but could also compromise 3D continuity, affecting landmarks that span multiple slices. This variability likely impacts the previous steps leading to difficulties in obtaining accurate landmark predictions across several slices. Additionally, the resampling process itself can introduce noise and artifacts into the data. The challenge lies in accurately detecting the center of these landmarks, especially when the heatmaps exhibit noise, multiple peaks, or spread-out values. In Figure 7.5, examples of acceptable and unsatisfactory predictions are displayed.



(a)



(b)

Figure 7.5: Comparison of predictions using BatchNorm. (a) Good prediction on landmark A5. (b) Bad prediction on landmark A9.

7.4.4 5-fold results

Previously documented in Section 7.1, all the folds of the 5-fold were used for training and validation. Only one model was chosen for this comprehensive evaluation. For the DATASET_AXIAL, the simple 3D U-Net with BatchNorm, *unet3d*, was used. The results indicate variability across the folds, with MAE values ranging from 31.81 $\pm$ 17.72 mm to 54.12 $\pm$ 17.73 mm, with an average MAE of 42.34 $\pm$ 21.33 mm (Table 7.5). The standard deviation values are relatively high, suggesting that the model's performance varies across different subsets of the data.

Table 7.5: Average MAE (mm) value for 5-fold axial *unet3d* model (across all landmarks)

Fold	**MAE** Mean±Std (mm)
model_unet3d_fold_1	44.91 ± 20.21
model_unet3d_fold_2	31.81 ± 17.72
model_unet3d_fold_3	54.12 ± 17.73
model_unet3d_fold_4	39.67 ± 26.13
model_unet3d_fold_5	41.19 ± 24.85

Overall, the evaluation of DATASET_AXIAL did not meet expectations and highlighted the need for further tuning. Despite the disappointing results, the various experimentations revealed fundamental hyperparameters that are crucial for such approaches, such as Dice+CE loss and batch size adjustments. These parameters were essential for achieving initial fitting and establishing a good starting point. Despite these improvements, the models exhibited some degree of overfitting. The assessment of different techniques within the architectures showed promise; however, they did not consistently outperform simpler architectures, indicating a need for further refinement and study. Additionally, the evaluation highlighted the critical impact of preprocessing steps, such as resampling and ground truth construction on landmark detection accuracy. Overall, while the tuning process yielded valuable insights and some models showed potential, significant improvements are still necessary.

7.5 DATASET_SAGITTAL

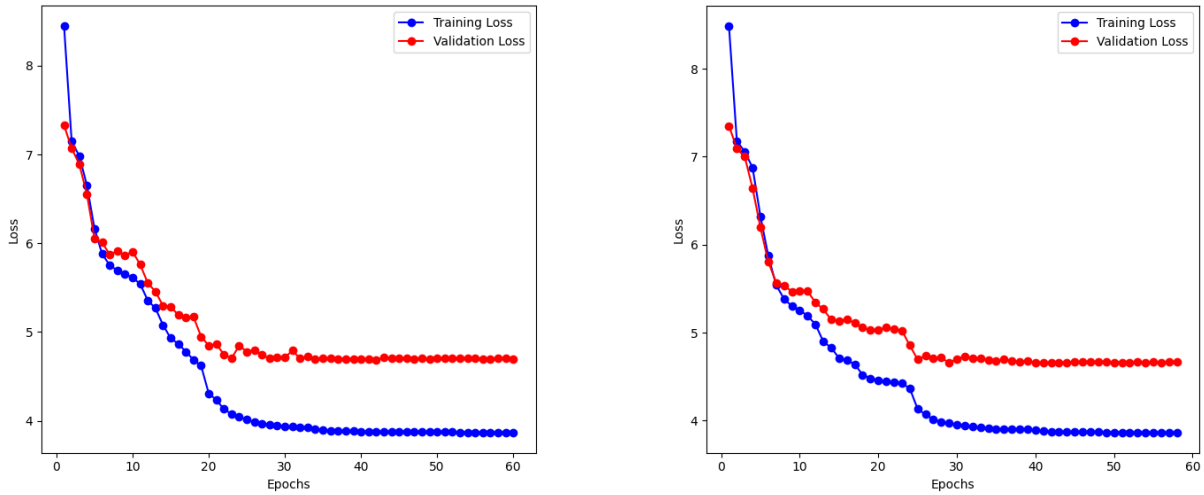
7.5.1 Training Approach

The training approach was very similar to that of the Axial subset. The best results were achieved using output channels and bridge values of [16, 32, 64, 128] and 512, which became the primary choice for most architectures; other values were tested but did not yield significant improvements. Given the volumes plane of aquisition of Sagittal, data augmentation techniques were filtered, with random horizontal flip not being selected given the clinical perspective. The probability of data augmentation was maintained like in the Axial subset. Hyperparameters tuned in the previous section were reused and tested, ensuring that the process did not start from scratch. Data augmentation techniques were carefully selected.

Regarding the learning rate study in the Sagittal subset, not many hyperparameters were tuned, it was merely a question of trying different values of out channels filters and *ReduceLROnPlateau* callbacks adjustments. The dropout regularization was applied more, especially in architectures where more calculations were performed, such as the ones where residual connections and attention gates were applied. The same loss function was applied to all the training done for the subset, it was the Dice+CE loss. The first model tested was the simple 3D U-Net with the respective output channel values for the encoder mentioned earlier, named *unet3d*. During the fine-tuning of this specific subset was performed through the increase of data augmentation probability, it increased the dynamic access to synthetic data and was benefic regarding the fitting of each model. Both BatchNorm and GroupNorm were tested and compared (Figure 7.6)

7.5.2 Loss

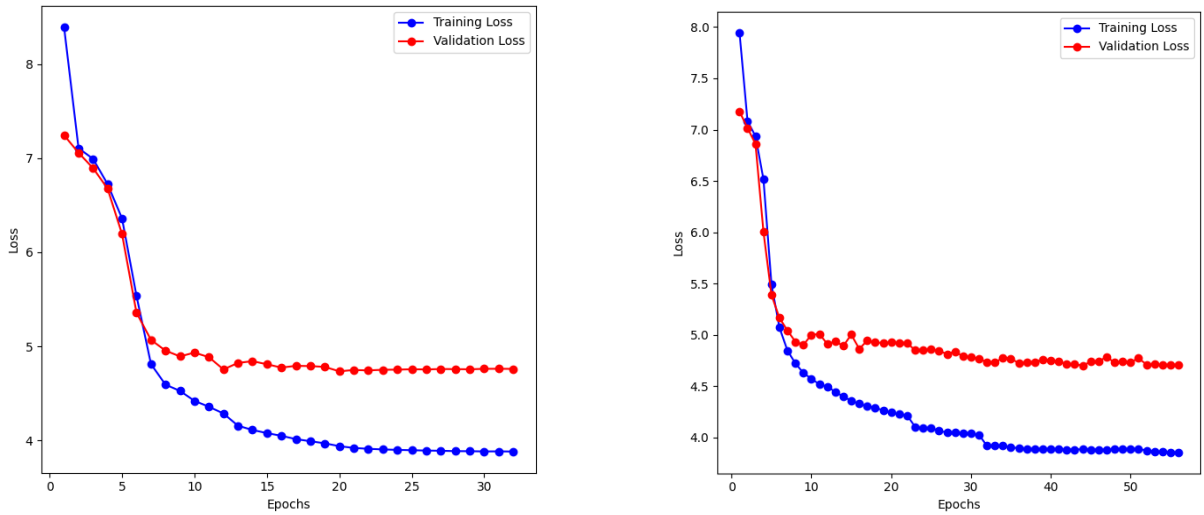
The number of initial epochs used for training were the same used for the previous subset, up to 100. Both normalization show signs of overfitting given that for about 25 epochs the validation loss did not fluctuate (Figure 7.6). This analysis suggested that the models would converge quicker that then DATASET_AXIAL's models. Although these are different subsets it was an interesting detail. Therefore the future fitting would be made with less epochs, in the range of 60 and 70.



(a) (b)

Figure 7.6: Loss learning curves for Simple 3D U-Net, unet3d: (a) BatchNorm, (b) GroupNorm.

The comparison of loss learning curves for the 3D U-Net with attention mechanisms, named *attunet3d* reveals distinct differences between BatchNorm and GroupNorm (Figure 7.7).



(a) (b)

Figure 7.7: Loss learning curves for 3D U-Net with attention mechanisms in the decoder, *attunet3d*: (a) BatchNorm, (b) GroupNorm.

BatchNorm exhibits a rapid initial decrease in both training and validation loss, with training loss converging around 4

and validation loss around 5 after 30 epochs. It shows a notable generalization gap, indicating potential overfitting. Similiarly, GroupNorm demonstrates a more gradual decline in loss, with the training loss stabilizing just below 4 and validation loss around 4.8 after 50 epochs. GroupNorm maintains a similiar generalization gap, with small oscillations between the same validation values. Thus, it is hard to highligth a specific normalization type as more effective in terms of stability and generalization. The main distinction lies in the number of epochs required for convergence.

7.5.3 MAE

The average MAE values were taken for each landmark, for BatchNorm (Figure 7.8) and GroupNorm (Figure 7.9). All measurements are in mm. In BatchNorm assessement, all models obtained very similar values. The Simple 3D U-Net,*unet3d_BN*, shows consistent performance, particularly excelling in channels S1, S4 and S5 with MAE values of 1.73 mm, 1.76 mm and 1.53 mm, respectively. The Residual U-Net with residual connections, *residualunet3d_BN*, while exhibiting higher variability, achieves its lowest MAE in channels S0 and S1, with 2.21 mm and 2.11 mm respectively, indicating some challenges in these landmarks. The only landmark where this model excels is S3 with 1.61 mm. The Attention U-Net *attunet3d_BN* demonstrates competitive results, especially in channels S3 and S5, 1.62 mm and 1.60 mm respectively. However, the Residual Attention U-Net *resattunet3d_BN* stands out with consistently low MAE values across most channels, particularly excelling in channels S5 with 1.52 mm and S6 with 1.27 mm, showcasing its robustness in accurate landmark detection.

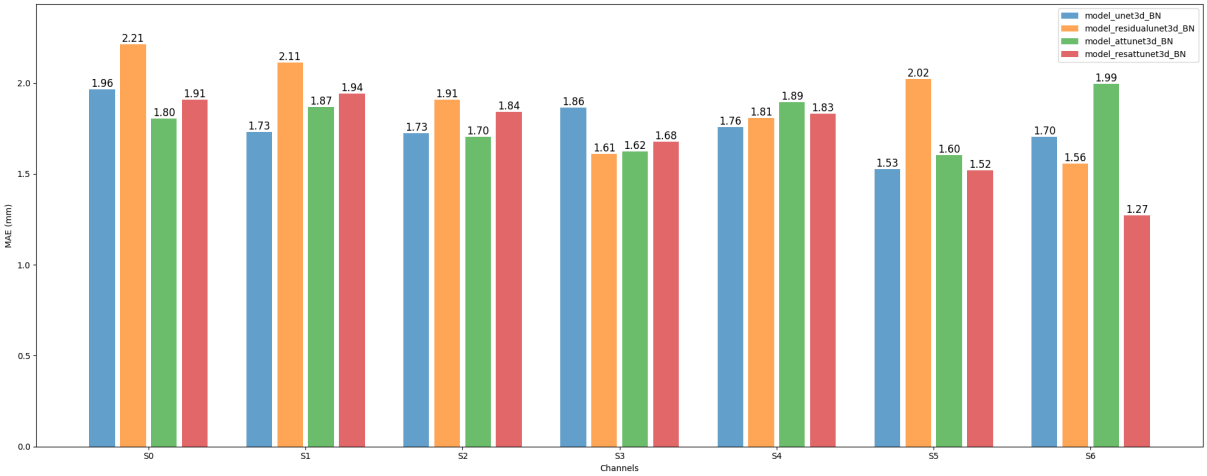


Figure 7.8: MAE for each of the architectures, for each of the landmarks in the Sagittal dataset, in mm, using BatchNorm.

Using GroupNorm (Figure 7.9), the *unet3d_GN* exhibits significant variability, with its highest MAE in channel S1 with 3.21 mm and S3 with 2.67 mm, but shows better performance in channels S5 and S6, with 1.59 mm and 1.27 mm respectively. The *residualunet3d_GN* demonstrates balanced performance, excelling in channels S2 achieving 1.66 mm and in S3 1.58 mm, though it struggles in channel S4 with 1.95 mm. The *attunet3d_GN* performs competitively across several channels, particularly in channels S1 and S6, with 1.53 mm and 1.34 mm respectively. The *resattunet3d_GN* shows less strong performance across most channels, struggling in channels S0 with 2.94 mm and S2 with 2.58 mm. Overall, the *resunet3d_GN* and *attunet3d_GN* models exhibit robust performance with lower MAE values across multiple channels using GroupNorm, making them reliable choices for accurate landmark detection in sagittal images.

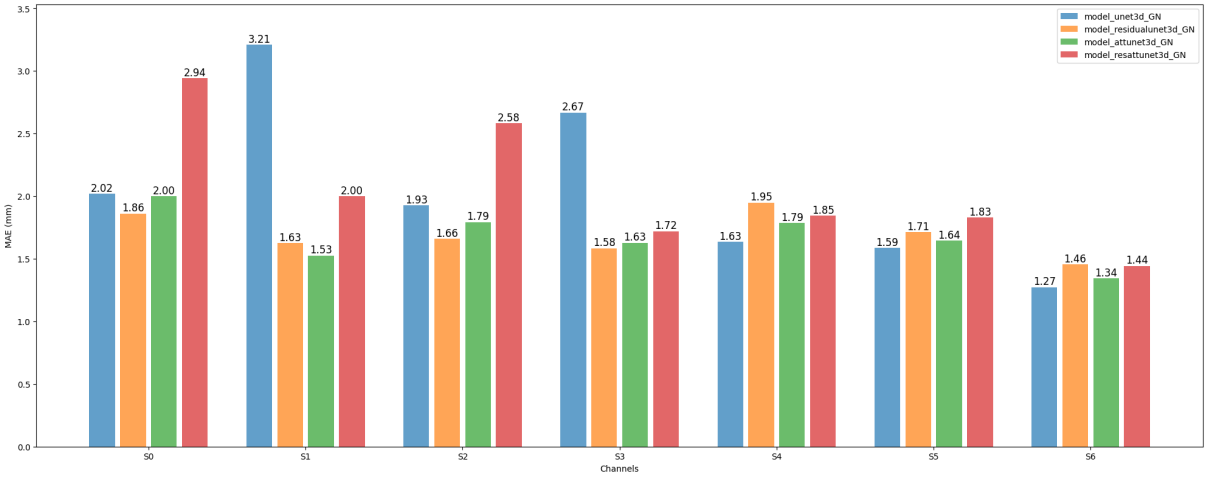


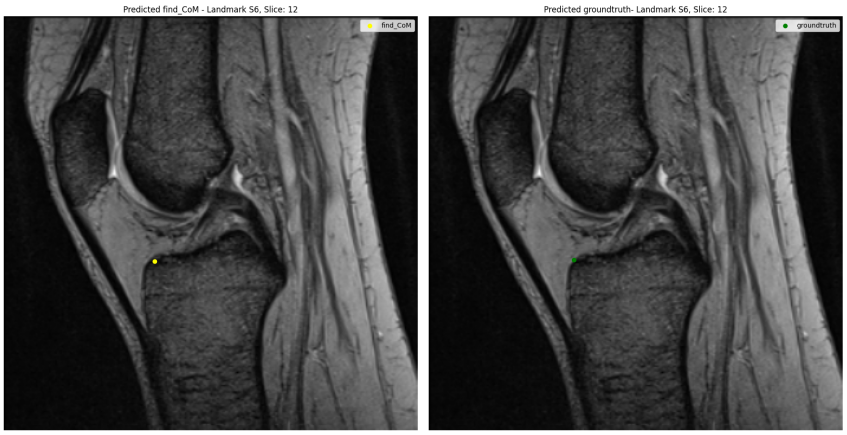
Figure 7.9: MAE for each of the architectures, for each of the landmarks in the Sagittal dataset, in mm, using GroupNorm.

Table 7.6 compares the average MAE values for the various sagittal models. The BatchNorm models generally exhibit lower MAE values with less variation, indicating more consistent performance across the architectures. Specifically, model_ResAttUnet_BN achieves the lowest MAE of 1.71 $\pm$ 0.23 mm. On the other hand, GroupNorm models show slightly higher MAE values, with model_residualunet3d_GN and model_attunet3d_GN performing comparably well at 1.69 $\pm$ 0.15 mm and 1.67 $\pm$ 0.24 mm, respectively. Evaluations of BatchNorm and GroupNorm revealed that while BatchNorm showed rapid convergence, it faced overfitting issues, whereas GroupNorm provided better stability and generalization, despite a slower convergence. This suggests that while BatchNorm provides more stable and lower error rates initially, certain GroupNorm models also offer competitive performance, particularly in terms of stability and generalization over time.

Table 7.6: Average MAE (mm) value for each sagittal model (across all landmarks) using BatchNorm and GroupNorm

Normalization	**Architecture**	**MAE** Mean$\pm$Std (mm)
BatchNorm	model_unet3d_BN	1.75 $\pm$ 0.13
	model_residualunet3d_BN	1.89 $\pm$ 0.23
	model_attunet3_BN	1.78 $\pm$ 0.14
	model_ResAttUnet_BN	1.71 $\pm$ 0.23
GroupNorm	model_unet3d_GN	2.05 $\pm$ 0.62
	model_residualunet3d_GN	1.69 $\pm$ 0.15
	model_attunet3d_GN	1.67 $\pm$ 0.24
	model_resattunet3d_GN	2.05 $\pm$ 0.49

In Figure 7.10, examples of acceptable and unsatisfactory predictions are displayed.



(a)

(b)

Figure 7.10: Comparison of predictions using BatchNorm: (a) good prediction on landmark S6, (b) bad prediction on landmark S5.

After assessing loss tuning, the assumption is that the subset achieved surprisingly good results. The base height and width for SAGITTAL and AXIAL images were (256,256). The SAGITTAL subset analysed in Section 6.1 showed that most original images had around 320 rows and columns. Resampling to 256 rows and columns represented a significant reduction, with some volumes having resolutions as high as 500-600 rows and columns, requiring more aggressive resampling. This resampling could result in loss of details and artifacts, potentially impacting model performance and increasing MAE values. However, the performance showed good results, for the particular fold. Further validation with a 5-fold evaluation was conducted providing a clearer demonstration of the subset's performance. In Figure 7.10, examples of acceptable and unsatisfactory predictions are displayed.

7.5.4 5-fold results

Previously documented in Section 7.1, all the folds of the 5-fold were used for training and validation. Only one model was chosen for this comprehensive evaluation. For the DATASET_SAGITTAL, the simple 3D U-Net with BatchNorm, *unet3d*, was used. The results indicate a high level of consistency across the folds, with MAE values ranging from 1.60 $\pm$ 0.21 mm to 1.77 $\pm$ 0.23 mm, with an average MAE of 1.65 $\pm$ 0.21 mm (Table 7.7). The standard deviation values are relatively low, suggesting that the model's performance is stable across different subsets of the data.

Table 7.7: Average MAE (mm) value for 5-fold sagittal *unet3d* model (across all landmarks) using BatchNorm

Fold ID	**MAE** Mean±Std (mm)
model_unet3d_fold_1	1.77 ± 0.23
model_unet3d_fold_2	1.69 ± 0.30
model_unet3d_fold_3	1.61 ± 0.21
model_unet3d_fold_4	1.60 ± 0.20
model_unet3d_fold_5	1.60 ± 0.13

The evaluation of DATASET_SAGITTAL showcased a comprehensive application of techniques and insights gained from the Axial subset analysis. The training approach adapted successfully to the unique challenges presented by the Sagittal dataset, maintaining consistency in hyperparameter choices and data augmentation techniques. Experimentation with the Dice+CE loss function continued to yield promising results. The 3D U-Net architecture with the chosen parameters showed reliability. The addition of dropout regularization and data augmentation enhanced model fitting. Overall, MAE evaluations highlighted the competitive performance of the models, showing strong and consistent results. The resampling process, despite reducing image resolutions significantly, did not adversely affect model performance. The 5-fold cross-validation further confirmed the stability and reliability of the *unet3d* model, evidenced by low std values across different data test subsets. These results provide an excellent impact over the goal to effectively detect landmarks for the PFI evaluation.

7.6 DATASET_DYNAMIC

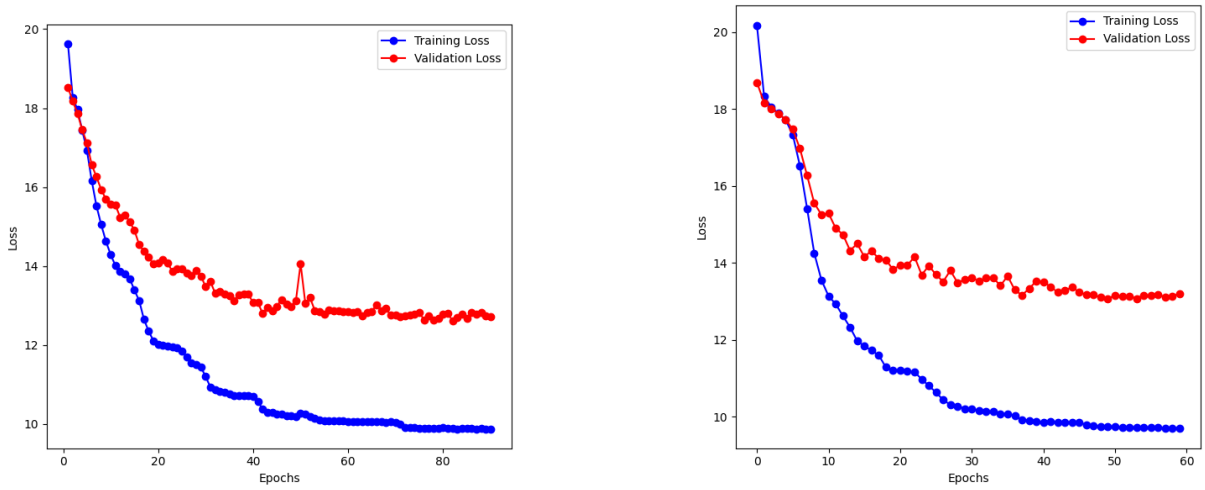
As from the previous development for the Sagittal and Axial, the knowledge were migrated to the construction of Dynamic results. Given the acquisition plane and the number of landmarks marked on the data, this was the heaviest subset in input terms. The DATASET_DYNAMIC was the only subset in which cross-validation wasn't performed due to practical constraints. The reason behind this was the limited availability of Computer 2. When attempts were made to run the cross-validation process, the computer consistently crashed—not due to OOM errors, but because the power source of the computer itself failed. This hardware limitation made it impractical to perform a full 5-fold cross-validation on the dynamic subset on the available amount of time.

7.6.1 Training Approach

The tuning process for this subset was the shortest, with the entire process conducted on Computer 2. Hyperparameters tuned in previous Axial and Sagittal subsets were reused and tested. Despite the relatively larger training data, the batch size remained at 2. Dropout regularization technique was used more frequently on *unet3d*, *resunet3d* and *attunet3d*. The normalization group values were higher, up to 16. In addition to the number of volumes for fitting, this subset had the highest output channels at 19, requiring significantly more processing power and consequently more time to fit the models.

7.6.2 Loss

The validation loss assessment was conducted similarly to the previous two subsets, using the same loss function, Dice+CE, for all training. As usual, training began with the use of the simple 3D U-Net, *unet3d*. Although the subset had higher availability data augmentation was to applied with a lower probability ranging 50% to 60%, and with lower number of epochs. Through the tuning process the number of epochs increased, given that initial trainings sessions were using all of the epochs specified. Using a number of epochs range between 90 and 100 epochs decreasing down to 60 and 70, when analysing the learning curves the assumptions and factors were the same. The callbacks were reused, using an *EarlyStopping* patience of 14 and *ReduceLROnPlateau* patience of 8 with a reduction factor of 0.2 the validation loss was evaluated for *unet3d* (Figure 7.11), for both BatchNorm and GroupNorm.



(a) (b)

Figure 7.11: Loss learning curves for Simple 3D U-Net, *unet3d*: (a) BatchNorm, (b) GroupNorm.

In Figure 7.11a the training loss steadily declines from above 20 to just over 10 by the end of the training period (around 80 epochs), indicating effective learning from the training data. In contrast, the validation loss drops from approximately 18 to 14 within the first 20 epochs but then exhibits fluctuations and occasional spikes, stabilizing around 13 for the remainder of the training period. Overall, while the model effectively minimizes the training loss, the fluctuations in the validation loss suggest some sensitivity to the validation data. Despite these fluctuations, the validation loss stabilizes at a lower value than its initial state, indicating reasonable generalization capability. Considering that perhaps the model training suffered from overfitting, further tuning would likely be needed to improve validation stability. The number of epochs used wasn't the optimal as yet.

With a lower number of epochs, 60 the training loss using GroupNorm (Figure 7.11b) steadily decreases to just below 10, while the validation loss plateaus around 13 close to 14 after more agressive fluctuations. Compared to a BatchNorm model with a higher number of epochs, BatchNorm exhibited indicated potential overfitting however the validation loss reached lower values.

Attention 3D U-Net, *attunet3d*, with the same hyperparameters demonstrated similar validation loss final values (Figure 7.12). This training session used GroupNorm with 100 epochs, the behaviour remained similar to *unet3d* with BatchNorm. In this specific session, the callbacks were using an *EarlyStopping* patience of 16 and *ReduceLROnPlateau* patience of 10 with a reduction factor of 0.2.

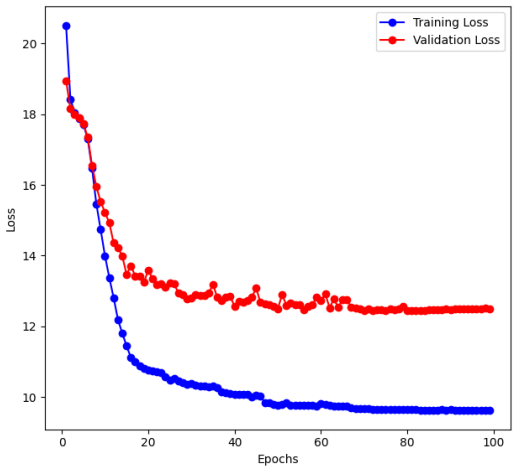


Figure 7.12: Loss learning curves for Simple 3D U-Net with attention mechanisms, *attunet3d* using GroupNorm.

Digging deeper, the learning curve demonstrates that training loss decreases rapidly from above 20 to below 10 within the first 40 epochs, indicating the model's quick adaptation to the training data. In contrast, the validation loss initially drops from approximately 20 to 11 within the first 20 epochs but then fluctuates around this value, with minor increases and decreases throughout the remaining epochs. This curve suggests reasonable generalization capability, although the fluctuations indicate some sensitivity to the validation data and potential overfitting. The adjusted callbacks, with patiences of 16 and 10 above, justify the tendency to overfit with this number of epochs. These tests were fundamental in assessing the optimal number of epochs for training across the various architectures.

Moving on the learning curve for the *resunet3d* model (Figure 7.13) showed a rapid initial decrease in training loss from around 20 to below 14 within the first 21 to 22 epochs, indicating quick learning. The validation loss also decreases rapidly initially but stabilizes close to 14. Following this, the training loss continues to decline steadily, eventually converging around 10, while the validation loss remains plateaued with minor oscillations. Assessing the *resunet3d* model demonstrated effective initial learning, however it faced challenges in improving validation performance, necessitating potential adjustments in regularization, higher amount of data augmentation and *EarlyStopping*'s callback adjustments .

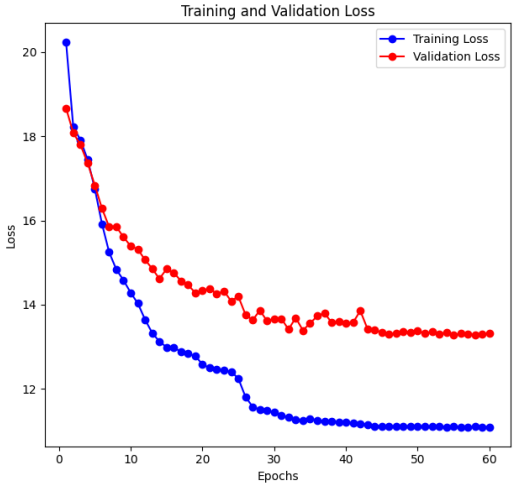
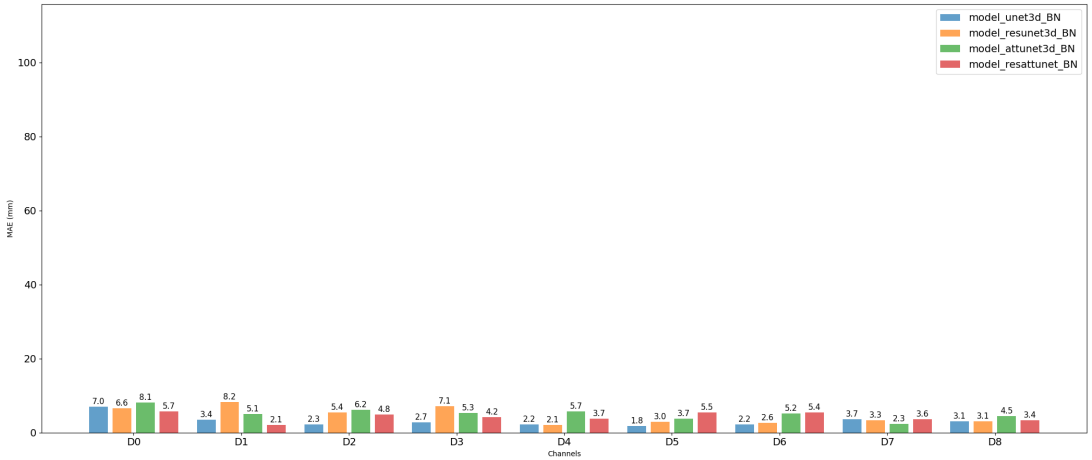


Figure 7.13: Loss learning curves for Simple 3D U-Net with residual connections, *resunet3d* using BatchNorm.

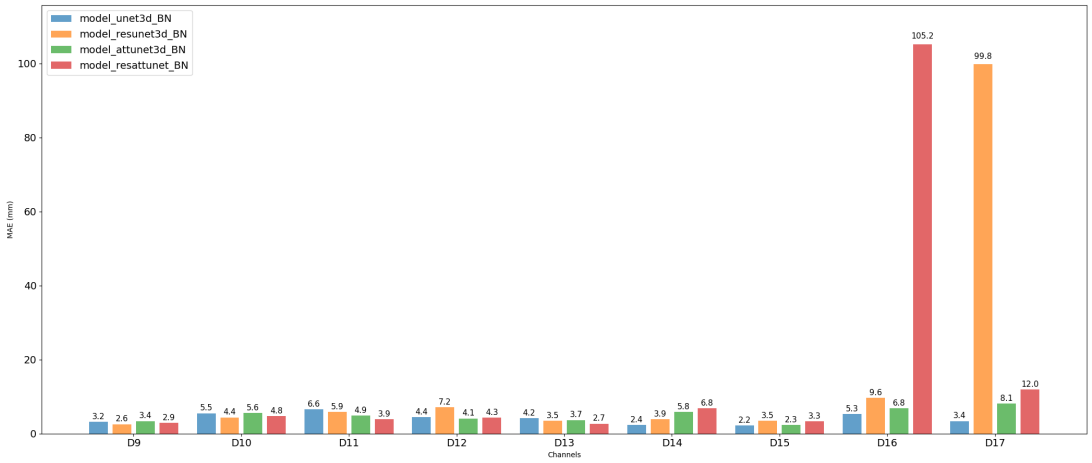
The tuning process overall revealed that, despite effective training loss minimization for the various architectures, validation loss exhibited fluctuations, suggesting sensitivity to the validation data and overfitting signs. This highlights that regardless of the availability of subsets, data augmentation and other strategies would further improve the training.

7.6.3 MAE

The average MAE values were taken for each landmark, both for the BatchNorm (Figure 7.14) and for the GroupNorm (Figure 7.15) across the tuned models. All measurements are in units of mm.



(a)

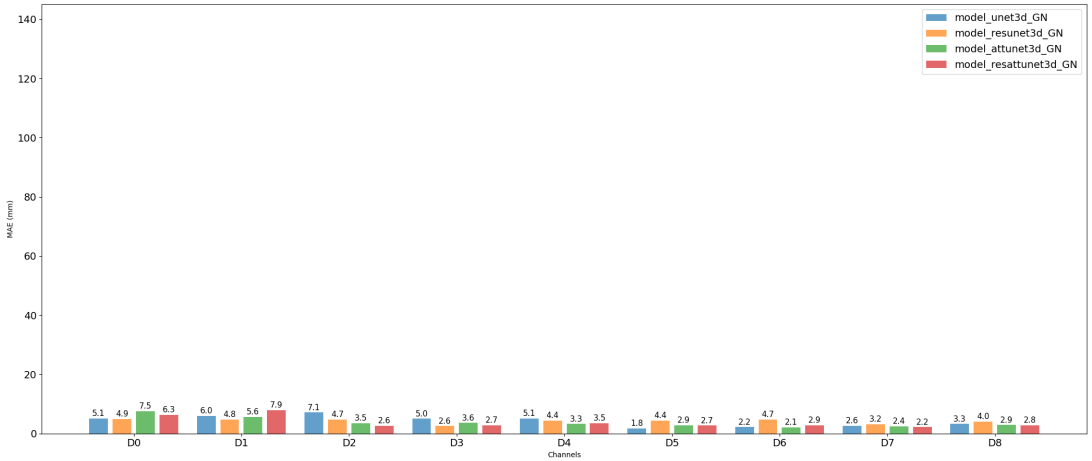


(b)

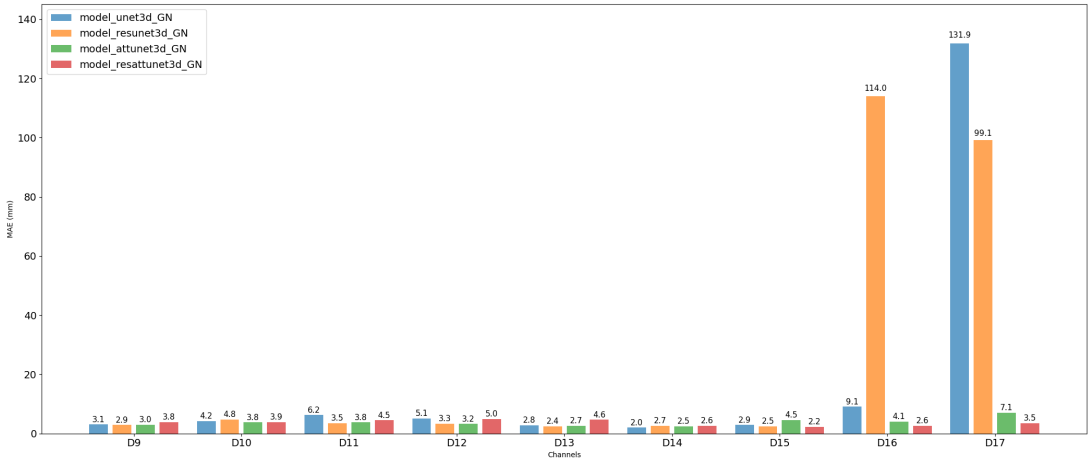
Figure 7.14: MAE for each of the architectures, for each of the landmarks in the Dynamic dataset, in mm, using BatchNorm: (a) landmark D0 to D8, (b) landmark D9 to D17.

The Simple U-Net with BatchNorm, *unet3d_BN* generally shows competitive MAE values, excelling in channels D1, D2, D3, D4, D5, D6, D8, D14 and D17. For instance, in channels D5 and D15, *unet3d_BN* records an MAE of 1.8 and 2.2 mm respectively, markedly lower than others. In comparison, the *resunet3d_BN* model performs well across several channels, for instance D4, D8, and D9, but shows the highest MAE in channel D17, with 99.8 mm. The *resnet3d_BN* and *resattunet3d_BN* models are the less consistent ones, given the difficulty of finding both D17 and D16 landmarks. This particular outliers are interesting because the models where it happens apply residual connections. On MAE values using BatchNorm, the model with

better performance was *unet3d*, while *resunet3d* and *resattunet3d* are jeopardised given the last two landmarks outliers.



(a)



(b)

Figure 7.15: MAE for each of the architectures, for each of the landmarks in the Dynamic dataset, in mm, using GroupNorm: (a) landmark D0 to D8, (b) landmark D9 to D17.

For GroupNorm, *unet3d_GN* often records the lowest MAE in channels D5, and D14, with 1.8 mm and 2.0 mm respectively. However, the values get worse for example, in channels D16 with 9.1 mm, and D17 with 131.9 mm. This latter landmark result was not effectively 'learned' by the model. The *resunet3d_GN* shows strong performance in various channels including D1 with 4.8 mm , D5 2.9 mm, D9 2.9 mm, and D15 with 2.5 mm, but has the highest MAE in channel D16 with 97.8 mm. High values like this demonstrate that the last two landmarks were hard to find.

The *attunet3d_GN* and *resattunet3d_GN* models display lower variability, with lower MAE values in several channels, indicating a more consistent performance compared to simple U-Net and the residual models. Both models perform similarly in channels D4, D5, D6 and D8, with attunet3d_GN still maintaining a slight edge in accuracy, with 3.3 mm, 2.9 mm, 2.1 mm and 2.9 mm respectively. Both models perform particularly similarly in channels D5, D10, D14.

The *resattunet3d_GN* shows lower error in the last landmarks channels D15 with 2.2 mm, D16 2.6 mm, and D17 with 3.5 compared to *attunet3d_GN*. This particular information is noteworthy given the application of attention mechanisms in these models, which enhance the focus on important details, such has these last landmarks. On MAE values using GroupNorm, the

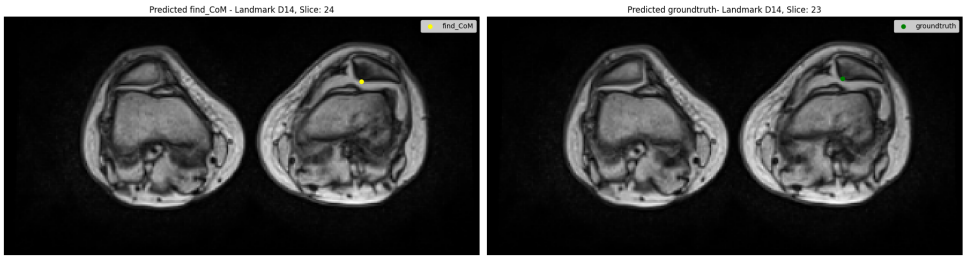
model with better performance was *resattunet3d*, while *unet3d* and *resunet3d* are jeopardised given the last two landmarks outliers.

From Table 7.8 models using BatchNorm generally showed better performance with lower MAE values compared to GroupNorm models. Specifically, *unet3d_BN* achieved an MAE of 3.65 mm $\pm$ 1.53 mm, while *attunet3d_BN* showed an MAE of 5.05 mm $\pm$ 1.62 mm. However, certain BatchNorm models (*resunet3d_BN* and *resattunet3d_BN*) were significantly affected by outliers in the last two landmarks, as indicated by their high standard deviations. In contrast, GroupNorm models initially presented higher MAE values; however, models like *attunet3d_GN* and *resattunet3d_GN* maintained low MAE values of 3.79 mm $\pm$ 1.49 mm and 3.69 mm $\pm$ 1.47 mm, respectively. It is worth noting that the Attention-based U-Net architecture consistently maintained lower error rates for this subset.

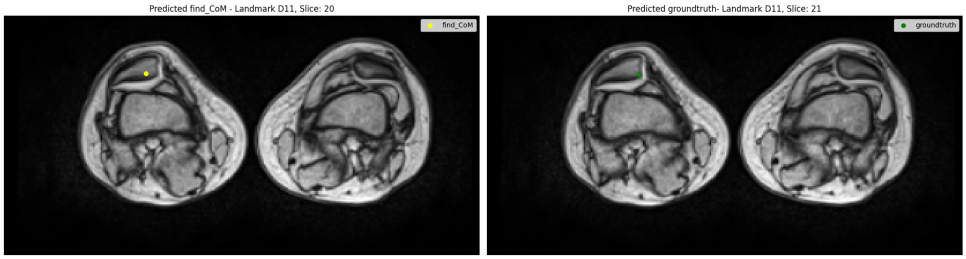
Table 7.8: Average MAE (mm) value for each model (across all landmarks) using BatchNorm and GroupNorm

Normalization	**Architecture**	**MAE** Mean$\pm$Std (mm)
BatchNorm	model_unet3d_BN	3.65 $\pm$ 1.53
	model_resunet3d_BN	10.09 $\pm$ 21.85
	model_attunet3d_BN	5.05 $\pm$ 1.62
	model_resattunet3d_BN	10.24 $\pm$ 23.13
GroupNorm	model_unet3d_GN	11.41 $\pm$ 29.29
	model_resunet3d_GN	15.16 $\pm$ 32.43
	model_attunet3d_GN	3.79 $\pm$ 1.49
	model_resattunet3d_GN	3.69 $\pm$ 1.47

In Figure 7.16, examples of acceptable and unsatisfactory predictions are displayed.



(a)



(b)

Figure 7.16: Comparison of predictions using BatchNorm: (a) good prediction on landmark D14, (b) bad prediction on landmark D11.

The evaluation of the Dynamic subset leveraged insights from previous subsets, applying them to a larger and more complex dataset. The MAE evaluations indicated competitive performance across models, although residual connection models struggled with specific landmarks, particularly the last two landmarks, highlighting areas for further refinement. Evaluating the subset using only one fold introduced a degree of uncertainty, underscoring the need for cross-validation. Implementing cross-validation would provide more robust and reliable results, ensuring consistent model performance across varying data subsets. However, this technique was not applied in the current evaluation.

7.7 Benchmarking

In order to compare our results with current state of the art studies, several specifications were considered. Relevant studies from Chapter 3 were selected based on their use of the same metric, MAE, to ensure a fair comparison. Out of the 6 studies found, the following were excluded: Danilov et al. (2022) [59] due to missing MAE values for test data, Caspersen et al. (2022) [72] as it only utilized MAE for optimal slice stage prediction and not for landmark detection, and Leitner et al. (2022) [77] which mentioned MAE but did not provide specific values. Table 7.9 displays specifications from other studies. It is important to note that these studies employed different architectures and focused on different clinical areas, with different image modalities.

Table 7.9: Benchmarking (NA: Not Available)

Study	**Clinical Area**	**Image Modality**	**Architecture(s)**	**MAE** Mean±Std (mm)
Liu et al. (2019) [44]	Pelvis	X-Ray	FR-DDH Network (Faster R-CNN elements with ResNet)	MAE (1.24 ± NA)
Jafari et al. (2022) [75]	Heart	US (echo videos)	U-LanD framework (Bayesian U-Net)	MAE (1.08 ± 0.89)
Shankar et al. (2022) [79]	Obstetrics	US	U-Net, Stacked Hourglass and HRNet	MAE (1.98 ± 0.89)
Our approach	Knee	MRI	3D U-Net	MAE (Axial: 42.34 ± 21.33, Sagittal: 1.65 ± 0.21, Dynamic: 3.65 ± 1.53)

In comparison with Liu et al. (2019) [44], which demonstrated robust performance with an MAE of 1.24 mm, our results were higher for the Axial subset with an MAE of 42.34 ± 21.33 mm, higher but closer for the Sagittal subset with an MAE of 1.65 ± 0.21 mm, and higher for the Dynamic subset with an MAE of 3.65 ± 1.53 mm. The dimensionality of the input data played a significant role. The authors leveraged the ResNet101 architecture with transfer learning pretrained on ImageNet, incorporating Faster R-CNN elements to efficiently adapt it for 2D landmark detection. Liu et al. (2019) also focused on a smaller number of critical landmarks, 6 for dysplasia of the hip diagnosis, whereas our approach involved a more extensive landmark detection process, with 11 landmarks in the Axial subset, 7 in the Sagittal subset, and 18 in the Dynamic subset. Additionally, the dataset variability was another important factor. The authors handled a dataset with a wide age range and various conditions, with higher availability adding to the transfer learning gain, which was useful for model generalization.

In comparison with Jafari et al. (2022) [75], which obtained an MAE of 1.08 ± 0.89 mm, our results were higher for the Axial subset with an MAE of 42.34 ± 21.33 mm, higher but closer for the Sagittal subset with an MAE of 1.65 ± 0.21 mm, and lower performance in the Dynamic subset with an MAE of 3.65 ± 1.53 mm. In this study, the dimensionality of the input data also played a significant role. The authors leveraged the BU-Net architecture, a Bayesian adaptation of the U-Net, incorporating uncertainty estimation techniques to adapt it for 2D landmark detection and provide better regularization. Our 3D nature of MRI image complexity may have contributed to the higher MAE values observed in some subsets, as opposed to the study's 2D US images. The study focused on detecting only 2 critical landmarks, whereas our approach involved a higher number of landmarks for all 3 subsets. Additionally, the study handled a dataset with echo videos of 4,493 patients, demonstrating good dataset availability.

In comparison with Shankar et al. (2022) [79], which demonstrated robust performance with an MAE of 1.98 ± 0.89 mm, our results were higher for the Axial subset with an MAE of 42.34 ± 21.33 mm, better for the Sagittal subset with an MAE of 1.65 ± 0.21 mm, and lower for the Dynamic subset with an MAE of 3.65 ± 1.53 mm. The study exploited multiple DL backbone networks, including the U-Net built for 2D, a Stacked Hourglass Network, and HRNet, to improve the accuracy of caliper placement for 2D US images. The authors focused on detecting 6 critical caliper landmarks, a lower number compared to any of our subsets. The study also benefited from a dataset with 1,192 images from US examinations.

Overall, the 3 studies used for benchmarking provided valuable insights. Each study focused on distinct clinical areas, different from the knee, and all followed a 2D approach, which likely simplified the complexity of landmark detection compared to our 3D approach. All the studies had access to large and comprehensive datasets, promoting generalizability in the models used. However, none of the studies explicitly describe the use of k-fold cross-validation, suggesting that their results may lack some robustness and reliability. The number of landmarks was also a crucial factor, with our subsets requiring the detection of a higher number of points, making it a more extensive and complex landmark detection process.

7.8 Conclusion

The evaluation process for the various datasets incorporated cross-validation to ensure robust model performance and minimize validation score variance. For the Axial and Sagittal subsets, a 5-fold cross-validation method was employed, dividing data into training and validation sets to fine-tune the models effectively. This approach used the same seed value across all subsets, maintaining consistent training and validation splits, using always the same fold for each training. The test datasets were carefully balanced to include healthy and pathological cases, as well as gender and laterality proportions, providing a comprehensive test scenario for each subset.

For the Axial subset, the evaluation highlighted the need for further tuning, as initial results were suboptimal despite identifying key hyperparameters like Dice+CE loss and batch size. Overfitting remained an issue, and preprocessing steps such as resampling and ground truth construction critically impacted landmark detection accuracy. Given the importance of this specific subset for PFI evaluation, particularly the points being detected, improvement is necessary. Future steps should focus on data preprocessing and ground truth adjustments, along with more detailed tuning of the models. A deeper study is required to analyse the outputs correctly and to understand why the results are currently inadequate for landmark detection.

The Sagittal subset yielded significant and promising results. These results proved to be very different from those of the Axial subset, providing useful insights for future PFI assessment. Consistent hyperparameter choices and data augmentation techniques were maintained. The use of Dice+CE loss and dropout regularization enhanced model fitting, with GroupNorm providing better stability and generalisation despite slower convergence. The 5-fold cross-validation confirmed the stability and reliability of the models, particularly the *unet3d* architecture, which showed strong performance across various test subsets. The Sagittal subset proved to be very close to effective detection, but further evaluation within a clinical context is necessary to validate these results.

The evaluation of the Dynamic subset applied previous insights to a larger and more complex dataset. With the correct batch size and extensive use of dropout regularization, the tuning process showed effective training loss minimization but validation loss fluctuations, indicating potential overfitting. The subset properties detailed in Section 6.1 and the number of landmarks being detected were sensitive factors that impacted the model. However, several models showed promising results. MAE evaluations revealed competitive performance across models, though residual connection models struggled with specific landmarks, particularly the last two. The use of attention mechanisms proved to be a good option for this data, particularly when comparing results with *attunet3d*. The absence of cross-validation in this subset introduced uncertainty, underscoring the need for its implementation to ensure more robust and reliable results across varying data subsets. With further work, these results would prove useful for PFI assessment.

The benchmarking section showed that, although our results were comparable to the state of the art levels demonstrated in the benchmark studies, the Axial subset exhibited some discrepancies. However, the Sagittal and Dynamic subsets produced results that align well with optimal state of the art outcomes, even surpassing one particular study [79] for the Sagittal subset. These findings provide satisfactory and promising outcomes, offering a positive indication for the task of normalizing and automating PFI diagnosis.

Chapter 8

Conclusions

This dissertation successfully developed DL-based models for landmark detection, crucial for the automatic and standardized diagnosis of PFI. By addressing the limitations of current diagnostic methods, such as the little consensus in the literature and the annotation complexity from these methods, the solution aimed for the standardization of the diagnostic process, enhancing precision in clinical settings. The comprehensive evaluation of these approaches, alongside the contributions in literature review, system development, and algorithm validation, underscored the potential of DL in advancing medical diagnostics. Before jumping into the development, it was required to understand the context on which the dissertation would fit.

Despite the significant progress made in applying DL to landmark detection in knee medical images, the journey towards fully automated, accurate and standardized diagnostics is ongoing. Future research will likely focus on refining the methodologies and techniques discussed. As researchers continue to push the boundaries of what DL can achieve in medical diagnostics, the ultimate goal remains clear: to provide clinicians with reliable tools that support early detection, accurate diagnosis, and effective treatment planning. Therefore, it is crucial to conduct a thorough review that addresses the existing literature on the area.

An overview of DL within the broader field of ML, focusing on its application to MIA, was conducted. It began by exploring DL concepts and learning paradigms and their relevance to various data types and modeling challenges. Key DL architectures, such as CNN, RNN and others, were discussed, highlighting their impact on DL tasks. The discussion then shifted to DL's role in MIA, tracing its evolution from rule-based systems to modern, automated, data-driven approaches. It emphasized the importance and challenges of applying DL in medical imaging, particularly for tasks such as object detection, segmentation, and registration. It also highlighted some of the current trends in the MIA environment, noting CNN-derived solutions such as U-Net, ResNet, and VGG. Emerging techniques using attention gates, transfer learning, and GANs provided a good visualization of the current context. In accordance with Chapter 1, the overview set the stage for a detailed discussion on the landmark detection task in medical images, particularly for guiding future research in this critical area.

A Review on DL approaches for Landmarks Detection on Medical images was conducted. The segment highlights the importance of data dimensionality and quality in anatomical landmark detection using DL algorithms. Several insights were assessed. U-Net was identified as a versatile architecture, particularly effective in cephalometric analysis, with only two studies referring to the knee area. Combining U-Net with ResNet and VGG models enhanced performance, especially in data-scarce environments. MRI modality was discovered in seven studies, with the major medical imaging modality being X-ray. Various pipelines for landmark detection were identified, with Direct Heatmap Probability Mapping, often combined with probabilistic and regression approaches, being a common method. Key metrics like ME, MAE, RMSE, and SDR are crucial for evaluating model accuracy and reliability. The review concluded that the optimal DL solution for landmark detection transcends a one-size-fits-all approach, emphasizing the importance of assessing various crucial factors to determine the best solution. Future

research should focus on analyzing the critical role of data dimensionality and quality, assessing the best architecture for the imaging modality, and integrating foundational models. Leveraging deep feature extraction and transfer learning can enhance model performance. The metrics used should be correctly chosen to evaluate model accuracy and reliability, reflecting the precision required in the specific clinical context. Building upon these insights and conclusions from the review, the next step was to develop a well-constructed outline strategy for the dissertation solution.

A well-constructed outline strategy was essential for developing the solution. Thus, a comprehensive analysis of how the solution would be constructed was necessary, and the CRISP-DM framework was selected as it provided a theoretical strategic method to guide this process. Each phase ensured a systematic approach. Business Understanding involved identifying key factors for a successful workflow, including acquiring knowledge and determining viable approaches, ultimately selecting one that would be feasible and effective to follow. The chosen approach focused on 3D segmentation of Gaussian heatmaps into points at the voxel level, utilizing a 3D U-Net architecture, chosen for its robustness in handling complex 3D data. Data Understanding assessed anatomical structures, data quality, and availability. Data Preparation created a labelled dataset, conducted analysis, and prepared data for modelling. In the Modelling phase, various DL architectures were developed and fine-tuned. The Evaluation phase implemented performance metrics and benchmarking. Although the Deployment stage was not reached, the best-performing model was intended for integration into *OrthoKnee*. The CRISP-DM guidelines helped pave the best pipeline to implement this approach, ensuring a structured and methodical development process. By having a delined structure in which the solution could be developed, construction of it followed.

The foundation of the implemented work lay in the meticulous creation of the ground truth dataset. A deep understanding of the initial data quality and structure was assessed, it guided the annotation of landmarks. The raw dataset comprised knee MRI scans from 95 patients, with its structure organized by acquisition planes and further categorized by individual patients and knee sides. 3 subsets, DATASET_AXIAL, DATASET_SAGITTAL, and DATASET_DYNAMIC were constructed for further annotation and processing. After an extensive study of PFI, its risk factors, and indexes used for diagnostics, the anatomical landmark annotation protocol followed. This protocol was standardized based on a review of existing indexes, ensuring that the annotations were clinically relevant and broadly applicable. This proved to be an important and complex task, requiring attention to detail. All subsets required extensive filtering and labeling to ensure data consistency and accuracy. This involved the development of a GUI tool, ImageLabelGUI, designed to facilitate precise annotation of anatomical landmarks across the three types of subsets: Axial, Sagittal, and Dynamic. The GUI incorporated functionalities such as data visualization, image navigation, and landmark annotation, allowing for accurate identification and documentation of each landmark. ImageLabelGUI played a crucial role in this process, guiding users through the annotation of specific landmarks for each subset. The tool enabled precise marking of 11 anatomical landmarks for axial sequences, 7 for sagittal sequences, and 18 for dynamic sequences, adhering to a standardized protocol covering the previously studied key indexes for PFI assessment. This stage was one of the most important in the entire thesis, laying a solid foundation for subsequent stages of the research and ensuring that the data used for model training and evaluation was of the highest possible standard. By having all subsets fully constructed and filtered an analysis on each subset's volume would be required.

The initial phase of our approach involved a comprehensive descriptive analysis of the dataset, focusing on data distribution, pattern identification, and data quality assessment. This stage also included the examination of associated metadata, which informed subsequent preprocessing decisions. For the DATASET_AXIAL subset, the analysis revealed highly standardized slice thickness and varied image widths. The slice thickness was consistently at 3.0 mm, with two distinct slice spacing protocols (3.3 mm and 3.9 mm). Image widths varied significantly with peaks around 300, 320, and 416 pixels, and the slice depth had

a high concentration around 30 slices. The DATASET_SAGITTAL subset showed consistent slice thickness (3.0 mm) with two distinct spacing protocols (3.6 mm and 3.75 mm) and a uniform image width, primarily around 320 pixels. The slice depth exhibited several peaks around 24, 25, and 26 slices. The uniformity in slice thickness and image width facilitates standardized preprocessing steps. In contrast, the DATASET_DYNAMIC subset exhibited significant variability in slice numbers per volume and dimensions. Peaks were observed at 44 and 72 slices, with rows and columns showing variability, but consistent spacing at 1.5 mm. The histogram of rows had three prominent peaks at approximately 160, 200, and 384 rows, while the columns had peaks around 256, 320, and 384. The pixel spacing showed distribution around 0.80 mm and 1.5 mm. These differences influenced further preprocessing and modeling stages, emphasizing the need for tailored approaches to handle each subset's unique characteristics.

Following this, the creation of corresponding heatmaps using Gaussian distributions to generate 3D masks was a critical step. This was essential for preparing the data for modeling and was facilitated by the use of documented landmarks and their respective stored coordinates (obtained through ImageLabelGUI). The creation of the background channel demonstrated a well-thought-out 3D approach, aligning with how the network would learn to find the landmarks and classify each voxel. Furthermore, subsequent preprocessing steps, including resampling and normalization, were tailored to maintain the integrity of the volumetric data while preparing it for advanced analytical processes. Although some resampling cases could result in information loss or the introduction of artifacts, the choice of static dimension shapes, determined through descriptive subset analysis, and the selection of TFRecords for data format were driven by the need to optimize storage, serialization, and computational efficiency.

The modeling phase was characterized by the implementation of advanced CNN's, designed to handle the complexities of 3D MRI data. Through this process, several trending technologies were implemented through the classical U-Net architecture. The models ranged from a simple 3D U-Net to more sophisticated architectures incorporating residual connections, attention mechanisms, and a combination of both. Dynamic data augmentation played a significant role in this phase, introducing variability into the training process and improving model generalization. Techniques such as random rotation, translation, horizontal flipping, gaussian blur, and noise injection were applied on-the-fly during training, ensuring that the models were exposed to a diverse range of scenarios. This approach not only enriched the dataset but also tried to mitigate the risk of overfitting, allowing the models to perform robustly on unseen data.

The evaluation process for the various datasets incorporated cross-validation to ensure robust model performance and minimize validation score variance. For the Axial and Sagittal subsets, a 5-fold cross-validation method was employed This approach used the same seed value across all subsets, maintaining consistent training and validation splits. This division between training, validation, and testing was performed by subject, ensuring that each subject only entered one of the three groups. This method was employed to avoid bias and ensure the independence of the datasets. By keeping the data for each subject exclusive to a single set, the risk of data leakage was minimized. The test datasets were carefully balanced to include healthy and pathological cases, as well as gender and laterality proportions, providing a comprehensive test scenario for each subset. The evaluation metrics were chosen based on their ability to accurately reflect model performance and address the specific challenges of the datasets. From the reviewed studies in Chapter 3, MAE was selected for its consistency in measuring prediction accuracy without error cancellation. Due to the clinical complexity and subjectivity involved, no other evaluation metrics were used. In addition to the evaluation metrics, different loss functions were utilized to optimize model training. These included CE Loss, Dice Loss, and a combination of both, tailored to enhance the overlap between predicted and ground truth masks while ensuring probabilistic predictions.

Based on the final results and conclusions for each subset, it is evident that a one-size-fits-all approach is not optimal. Each

subset exhibited unique characteristics and hyperparameters adjustments. For the Axial subset, the evaluation highlighted the need for further tuning, as initial results were suboptimal despite identifying key hyperparameters like Dice+CE loss and batch size. Overfitting remained an issue, as evidenced by the learning curves, and the MAE results demonstrated the difficulty of training and correctly fitting this subset. Previous 5-fold operation, among all the models, the simple *unet3d* achieved the best results with an MAE of 30.99 $\pm$ 20.34 mm, closely followed by *attunet3d* with an MAE of 33.84 $\pm$ 19.84 mm. Architectures with residual connections, such as *residualunet3d* and *resattunet3d*, performed the worst. When applying the 5-fold validation *unet3d* was used due to the best performance. Across its folds, fold 2 achieved the best result with 31.81 $\pm$ 17.72 mm. Further investigation revealed that preprocessing steps such as resampling and ground truth construction critically impacted landmark detection accuracy, necessitating additional tuning to improve performance. A better analysis of overfitting and hyperparameter adjustments is also needed to enhance model accuracy. In contrast, the Sagittal subset achieved the best results. Consistent hyperparameter choices and data augmentation techniques were maintained, with the use of Dice+CE loss and dropout regularization significantly enhancing model fitting. This subset allowed for a comparison between gradient normalization techniques. Although their normalization processes differed, the results were quite similar. However, GroupNorm provided better stability and higher generalization, exhibiting a lower difference between validation loss and training loss despite slower convergence. In terms of results, before the 5-fold validation, *attunet3d* achieved the best values considering both normalization types, with an MAE of 1.67 $\pm$ 0.24 mm. Given that the results did not have significant variations between them, *unet3d* was selected for the 5-fold validation, using BatchNorm. For fold 5, *unet3d* achieved an MAE of 1.60 $\pm$ 0.13 mm. The 5-fold cross-validation confirmed the stability and reliability of the models, particularly for the *unet3d* architecture, which showed strong performance across various test subsets. Finally, the evaluation of the Dynamic subset applied previous insights to a larger and more complex dataset. The tuning process showed effective training loss minimization but validation loss fluctuations, indicating potential overfitting. MAE evaluations revealed competitive performance across models, though residual connection models struggled with specific landmarks, particularly the last two. Overall without 5-fold validation, the best model overall, was the *resattunet3d* model, which achieved an MAE of 3.69 $\pm$ 1.47 mm using BatchNorm. This model outperformed others in terms of both lower MAE and standard deviation, indicating more consistent performance across the several landmarks. The *attunet3d* model also performed well with an MAE of 3.79 $\pm$ 1.49 mm with GroupNorm, making it a competitive choice. The absence of cross-validation in this subset introduced uncertainty, underscoring the need for its implementation to ensure more robust and reliable results across varying data subsets.

The benchmarking process compared our results with other studies, highlighting significant variability in the number of landmarks, imaging modalities, and CNN architectures used. Our results were comparable to state of the art performances, with the Sagittal and Dynamic subsets producing outcomes that align well with, and even surpass, optimal benchmarks in one study [79]. Despite some discrepancies in the Axial subset, the findings were satisfactory and promising, positively indicating progress in normalizing and automating PFI diagnosis.

In conclusion, the project revealed varied outcomes across all three subsets using multiple architectures, indicating significant potential for further enhancement in this medical image analysis task. The complexity of handling and processing 3D data added a unique dimension to the approach, compared to traditional 2D methods. The challenges in accurately localizing anatomical structures in a 3D context were considerable, impacting the precision and utility of the models. Alongside this final assessment, it is worth noting that the knee area, particularly concerning PFI diagnosis, does not have a high number of research papers. This project sets a new standard for future research in the field and specifically for PFI normalization. Given that PFI diagnosis is a time-consuming process, manually performed with significant intra- and interobserver variability, this

project highlights the potential of AI-based algorithms, such as DL algorithms, to bring remarkable progress in MIA. By creating automatic models to predict objective indices more quickly and accurately, these technologies can significantly assist radiologists, reducing variability and saving substantial time in the joint assessment procedure. This role should galvanize the general standardization of PFI diagnosis and serve as a fundamental reference for subsequent studies, addressing current limitations and filling the literature gap.

The dissertation allowed to answer the RQs appointed in Chapter 1:

RQ 1 : What factors should be considered for anatomical landmarks detection on medical images based in DL algorithms?

Chapter 3 answered this RQ. Landmark detection in medical images is a complex task. Thus, several factors must be considered, such as imaging modalities, architecture choices, and performance evaluation metrics. Therefore, the optimal DL solution for landmark detection is not limited to a single approach. It requires a sophisticated analysis of the clinical context in which the task is performed, data attributes, the consideration for different methods and pipelines, the selection of architectures tailored to specific tasks, and the employment of sophisticated performance metrics evaluation. In addition to advances in the field, the integration of emerging models along with innovative approaches to data augmentation and transfer learning will be pivotal to overcoming the limitations posed by data scarcity and variability.

RQ 2 : Regarding the landmark labeling program, what are the key landmarks necessary for accurate PFI diagnosis, and how does the developed GUI tool support their precise annotation?

Chapter 5 answered this RQ. The landmark labeling program was essential for accurate PFI diagnosis, identifying key landmarks across several categories: trochlear dysplasia, patellar height, patellar lateralization, patellar tilt, and tibial tubercle lateralization. The developed GUI, ImageLabelGUI, facilitated precise annotation by providing an intuitive interface and standardized protocols. It supported data visualization, guided users through the annotation process, and ensured consistent and clinically relevant landmark identification. The tool's features included storing annotated landmarks and metadata in JSON and Excel formats, thus ensuring data integrity for future analysis. The dataset, comprising 993 sequences of knee MRI scans from 95 patients, was meticulously filtered and organized into axial, sagittal, and dynamic subsets. Each subset had specific instructions for marking landmarks, ensuring high-quality annotations and laying a robust foundation for accurate PFI diagnosis.

RQ 3 : What are the essential components and how should an automated pipeline be constructed for accurate and reliable landmark detection in MRI scans used in PFI diagnostics?

Chapter 6 answered this RQ. Firstly, it is essential to understand and manage the data. Essential data management tasks include annotating the MRI sequences to create accurate ground truth landmarks (through the use of a customised application for marking anatomical landmarks, such as ImageLabelGUI), facilitating the storage of metadata, and ensuring that the datasets are well-organized and anonymous. Second, data preprocessing is crucial, starting with descriptive analysis to understand the dataset's structure and characteristics, such as slice thickness, pixel spacing, and number of slices. Consistent resampling of the MRI volumes is also necessary to standardize the input dimensions for modeling. Techniques such as Gaussian heatmaps are used to construct ground truth masks that help convert the task into a heatmap regression problem. The pipeline should also

use advanced data formats, such as NIfTI and/or TFRecords, to efficiently handle large datasets and facilitate integration with TensorFlow for model training. Data augmentation techniques, such as random rotations, flips, translations, gaussian blurring, and noise injection, increase the dataset's diversity, and improve the model's generalization capabilities.

Finally, it is essential to use sophisticated DL models. The pipeline should advantage of architectures such as 3D U-Net, Residual 3D U-Net, and Attention U-Net adapted to 3D medical imaging tasks. These models should incorporate GroupNorm and/or BatchNorm normalization techniques, advanced loss functions that combine Cross-Entropy and Dice losses for optimal performance. Dynamic data augmentation and fine-tuning strategies, supported by callbacks for early stopping, learning rate adjustments, and model checkpoints to ensure robust and efficient training. In conclusion, constructing an automated pipeline for MRI-based landmark detection in PFI requires meticulous data management, comprehensive preprocessing, accurate ground truth construction, and the application of advanced DL models, all guided by a structured framework such as CRISP-DM.

RQ 4 : Do DL algorithms actually bring benefits to the PFI diagnostic process and what are the critical factors influencing their performance?

Chapter 7 answered this RQ.

DL algorithms indeed bring significant benefits to the PFI diagnostic process. One of the primary advantages is the substantial reduction in diagnosis time. Typically, diagnosing PFI through manual landmark detection in knee MRI scans can take around 20 to 30 minutes. However, with the implementation of DL algorithms, this time can be cut down to mere seconds, ranging from 3 to 10 seconds, depending on the complexity and volume of the data. This significant reduction in diagnosis time enhances efficiency for radiologists and technicians, allowing them to allocate more time to other critical tasks. It also facilitates quicker decision-making and treatment planning by doctors, improving patient care. For patients, a rapid diagnosis means shorter waiting times and quicker access to the necessary treatments. This efficiency can avoid the need for additional visits to the hospital, reducing travel costs and the associated inconvenience, thus increasing overall patient satisfaction. Hospitals and clinical centres benefit from higher throughput and more standardised diagnostic processes, which help to reduce intra- and inter-observer variability. In addition, there are substantial cost savings, as the dependence on technicians and radiologists for manual diagnosis decreases, allowing these professionals to focus on other essential tasks.

The performance of DL algorithms is influenced by several critical factors. Data quality is paramount; high-quality, well-labeled datasets are essential for training effective models. Hyperparameter tuning is crucial for optimizing model performance, and techniques like cross-validation play a significant role in ensuring robustness and generalizability. Normalization techniques, such as BatchNorm and GroupNorm, impact the stability and convergence of the models. Proper preprocessing steps, including resampling and accurate ground truth construction, are vital for maintaining the integrity and utility of the data. Furthermore, incorporating data augmentation techniques can help mitigate overfitting and enhance model robustness. Overall, these factors highlight the importance of a meticulous approach to model development and validation, ensuring that DL algorithms can provide reliable and efficient support in medical diagnostics.

8.1 Future work

The results presented in the previous chapters suggest significant opportunities to advance the goals of the project.

1. **Implementation of K-fold Cross-Validation for the Dynamic Subset**:

   Implementing k-fold cross-validation for the Dynamic subset is proposed to better evaluate the DL models with the specific acquisition plane and for a robust PFI index evaluation. This will ensure a more thorough assessment of the models' performance and generalization capabilities.

2. **Refinement of Gaussian Mask's Sigma Values**:

   Future work should explore refining Gaussian mask's sigma values by testing different lower or higher sigma values to evaluate their impact on model performance. This will help in identifying the optimal sigma value that enhances the precision of landmark detection.

3. **Utilization of Transfer Learning**:

   The use of transfer learning is expected to accelerate the training process and improve model performance by adapting pre-trained models to similar tasks. Examples mentioned in Chapter 3 include using a 3D ResNet backbone pre-trained on comprehensive video datasets [55] or initializing and fine-tuning the initial layers of multi-branch CNN architectures on pre-trained models like VGG19 [67]. This approach can significantly enhance model accuracy and reduce training time.

4. **Enhancement of Segmentation Accuracy with Hourglass Regression Networks**:

   Techniques such as Hourglass Regression Networks will be employed to refine the segmentation accuracy of heatmaps produced by future models. These networks, with their symmetric architecture that processes features at multiple scales, will allow for more precise localization and delineation of anatomical features in medical images.

5. **Integration into the OrthoKnee Application**:

   Future work involves fully integrating these algorithms into OrthoKnee and studying its usability after integration, aiming to automate the process of landmark detection. Additionally, incorporating tools like the constructed ImageLabelGUI will enable more efficient coordinate acquisition, which is fundamental for the ground truth process, thereby enhancing the application's overall usability and accuracy.

6. **Validation Protocol with End-Users**:

   A validation protocol involving end-users, such as technicians, radiologists, and surgeons in a hospital environment, will be developed. This will increase the number of diagnostic sessions of knee MRI images with both healthy and pathological subjects, providing a more comprehensive evaluation of the application's effectiveness in clinical settings.

7. **Human-in-the-Loop Approach**:

   An important aspect to consider is the human-in-the-loop approach, integrating human expertise into the DL pipeline during the annotation process to ensure high-quality ground truth data. Experts can review and correct automated annotations, improving the training data's quality and the model's accuracy. Additionally, during the model's deployment in clinical settings, the user can interact with the system to validate and refine the detected landmarks, perhaps with the help of reinforcement learning techniques, providing continuous feedback to improve the model over time.

8. **Dissemination and Collaboration**:

   Disseminating the achieved results through peer-reviewed ISI/Scopus journals and engaging in communication activities with hospitals and care facilities is crucial. These efforts aim to motivate interest and collaboration in end-user validation studies, bridging the gap between theoretical research and practical, clinical applications.

Bibliography

[1] Kai-Jonathan Maas, Malte Lennart Warncke, Miriam Leiderer, Matthias Krause, Tobias Dust, Jannik Frings, Karl-Heinz Frosch, Gerhard Adam, and Frank Oliver Gerhard Henes. Diagnostic imaging of patellofemoral instability. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 193(09):1019–1033, March 2021.

[2] Collin Krebs, Meaghan Tranovich, Kyle Andrews, and Nabil Ebraheim. The medial patellofemoral ligament: Review of the literature. *Journal of Orthopaedics*, 15(2):596–599, June 2018.

[3] Aishwarya Gulati, Christopher McElrath, Vibhor Wadhwa, Jay P Shah, and Avneesh Chhabra. Current clinical, radiological and treatment perspectives of patellofemoral pain syndrome. *The British Journal of Radiology*, page 20170456, January 2018.

[4] John E. Nolan, Patrick C. Schottel, and Nathan K. Endres. Trochleoplasty: Indications and technique. *Current Reviews in Musculoskeletal Medicine*, 11(2):231–240, May 2018.

[5] David H. Dejour, Guillaume Mesnard, and Edoardo Giovannetti de Sanctis. Updated treatment guidelines for patellar instability: "un menu à la carte". *Journal of Experimental Orthopaedics*, 8(1), November 2021.

[6] Federica Kiyomi Ciliberti, Lorena Guerrini, Arnar Evgeni Gunnarsson, Marco Recenti, Deborah Jacob, Vincenzo Cangiano, Yonatan Afework Tesfahunegn, Anna Sigríður Islind, Francesco Tortorella, Mariella Tsirilaki, Halldór Jónsson, Paolo Gargiulo, and Romain Aubonnet. CT- and MRI-based 3d reconstruction of knee joint to assess cartilage and bone. *Diagnostics*, 12(2):279, January 2022.

[7] Fleur V. Verhulst, Jordy D. P. van Sambeeck, Geerte S. Olthuis, Jasper van der Ree, and Sander Koëter. Patellar height measurements: Insall–salvati ratio is most reliable method. *Knee Surgery, Sports Traumatology, Arthroscopy*, 28(3):869–875, May 2019.

[8] Grant Buchanan, LeeAnne Torres, Brian Czarkowski, and Charles E Giangarra. Current concepts in the treatment of gross patellofemoral instability. *International journal of sports physical therapy*, 11(6):867—876, December 2016.

[9] Qin Ye, Taihen Yu, Yinbo Wu, Xiaonan Ding, and Xiangyang Gong. Patellar instability: the reliability of magnetic resonance imaging measurement parameters. *BMC Musculoskeletal Disorders*, 20(1), July 2019.

[10] Anne M. L. Meesters, Kaj ten Duis, Hester Banierink, Vincent M. A. Stirler, Philip C. R. Wouters, Joep Kraeima, Jean-Paul P. M. de Vries, Max J. H. Witjes, and Frank F. A. IJpma. What are the interobserver and intraobserver variability of gap and stepoff measurements in acetabular fractures? *Clinical Orthopaedics &amp Related Research*, 478(12):2801–2808, July 2020.

[11] Roberto M. Barbosa, Manuel Vieira da Silva, Carlos Sampaio Macedo, and Cristina P. Santos. Imaging evaluation of patellofemoral joint instability: a review. *Knee Surgery amp; Related Research*, 35(1), mar 2023.

[12] Le Lu, Yefeng Zheng, G. Carneiro, and Lin Yang. Deep learning and convolutional neural networks for medical image computing. In *Advances in Computer Vision and Pattern Recognition*, 2017.

[13] Meghavi Rana and Megha Bhushan. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*, 82(17):26731–26769, December 2022.

[14] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel S. W. Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, 4(1), April 2021.

[15] Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in Oncology*, 11, March 2021.

[16] Jun Zhang, Mingxia Liu, and Dinggang Shen. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions on Image Processing*, 26(10):4753–4764, October 2017.

[17] Toby O. Smith, Allan Clark, Sophia Neda, Elizabeth A. Arendt, William R. Post, Ronald P. Grelsamer, David Dejour, Karl Fredrik Almqvist, and Simon T. Donell. The intra- and inter-observer reliability of the physical examination methods used to assess patients with patellofemoral joint instability. *The Knee*, 19(4):404–410, August 2012.

[18] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.

[19] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9389, 2018.

[20] David Baur, Katharina Kroboth, Christoph-Eckhard Heyde, and Anna Voelker. Convolutional neural networks in spinal magnetic resonance imaging: A systematic review. *World Neurosurgery*, 166:60–70, 2022.

[21] Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: A review, 2020.

[22] Shankey Garg and Pradeep Singh. State-of-the-art review of deep learning for medical image analysis. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, December 2020.

[23] Iqbal H. Sarker. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6), August 2021.

[24] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1), jun 2015.

[25] Youzi Xiao, Zhiqiang Tian, Jiachen Yu, Yinshu Zhang, Shuai Liu, Shaoyi Du, and Xuguang Lan. A review of object detection based on deep learning. *Multimedia Tools and Applications*, 79(33–34):23729–23791, June 2020.

[26] Corina S. Păsăreanu, Divya Gopinath, and Huafeng Yu. *Compositional Verification for Autonomous Systems with Deep Learning Components: White Paper*, page 187–197. Springer International Publishing, nov 2018.

[27] Michael Hillebrand, Mohsin Lakhani, and Roman Dumitrescu. A design methodology for deep reinforcement learning in autonomous systems. *Procedia Manufacturing*, 52:266–271, 2020.

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

[29] Shoaleh Shahidi, Ehsan Bahrampour, Elham Soltanimehr, Ali Zamani, Morteza Oshagh, Marzieh Moattari, and Alireza Mehdizadeh. The accuracy of a designed software for automated localization of craniofacial landmarks on cbct images. *BMC Medical Imaging*, 14(1), sep 2014.

[30] Falk Schwendicke, Akhilanand Chaurasia, Lubaina Arsiwala, Jae-Hong Lee, Karim Elhennawy, Paul-Georg Jost-Brinkmann, Flavio Demarco, and Joachim Krois. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. *Clinical Oral Investigations*, 25(7):4299–4309, may 2021.

[31] Jiangchang Xu, Bolun Zeng, Jan Egger, Chunliang Wang, Örjan Smedby, Xiaoyi Jiang, and Xiaojun Chen. A review on ai-based medical image computing in head and neck surgery. *Physics in Medicine amp; Biology*, 67(17):17TR01, aug 2022.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[33] Siddhesh Bangar. VGG-Net Architecture Explained — siddheshb008. https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f. [Accessed 26-03-2024].

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

[35] Siddhesh Bangar. Resnet Architecture Explained — siddheshb008. https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d. [Accessed 26-03-2024].

[36] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[37] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[38] Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li, and Dimitris Metaxas. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2015.

[39] Yefeng Zheng, David Liu, Bogdan Georgescu, Hien Nguyen, and Dorin Comaniciu. 3d deep learning for efficient and robust landmark detection in volumetric data. In *Lecture Notes in Computer Science*, pages 565–572. Springer International Publishing, 2015.

[40] Matthieu Le, Jesse Lieman-Sifry, Felix Lau, Sean Sall, Albert Hsiao, and Daniel Golden. Computationally efficient cardiac views projection using 3d convolutional neural networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 109–116. Springer International Publishing, 2017.

[41] Bart Liefers, Freerk G. Venhuizen, Thomas Theelen, Carel Hoyng, Bram van Ginneken, and Clara I. Sánchez. Fovea detection in optical coherence tomography using convolutional neural networks. In Martin A. Styner and Elsa D. Angelini, editors, *SPIE Proceedings*. SPIE, February 2017.

[42] Yuanwei Li, Amir Alansary, Juan J. Cerrolaza, Bishesh Khanal, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Fast multiple landmark localisation using a patch-based iterative network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 563–571. Springer International Publishing, 2018.

[43] E.N.D. Goutham, Srikanth Vasamsetti, P.V.V. Kishore, and H.K. Sardana. Automatic localization of landmarks in cephalometric images via modified u-net. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, jul 2019.

[44] Chuanbin Liu, Hongtao Xie, Sicheng Zhang, Jingyuan Xu, Jun Sun, and Yongdong Zhang. Misshapen pelvis landmark detection by spatial local correlation mining for diagnosing developmental dysplasia of the hip. In *Lecture Notes in Computer Science*, pages 441–449. Springer International Publishing, 2019.

[45] Jiahong Qian, Ming Cheng, Yubo Tao, Jun Lin, and Hai Lin. CephaNet: An improved faster r-CNN for cephalometric landmark detection. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, April 2019.

[46] Aleksei Tiulpin, Iaroslav Melekhov, and Simo Saarakkala. KNEEL: Knee anatomical landmark localization using hourglass networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, October 2019.

[47] Junshen Xu, Molin Zhang, Esra Abaci Turk, Larry Zhang, P. Ellen Grant, Kui Ying, Polina Golland, and Elfar Adalsteinsson. Fetal pose estimation in volumetric MRI using a 3d convolution neural network. In *Lecture Notes in Computer Science*, pages 403–410. Springer International Publishing, 2019.

[48] Mohammad Eslami, Christiane Neuschaefer-Rube, and Antoine Serrurier. Automatic vocal tract landmark localization from midsagittal MRI data. *Scientific Reports*, 10(1), January 2020.

[49] Chuanbin Liu, Hongtao Xie, Sicheng Zhang, Zhendong Mao, Jun Sun, and Yongdong Zhang. Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip. *IEEE Transactions on Medical Imaging*, 39(12):3944–3954, December 2020.

[50] Tianyu Ma, Ajay Gupta, and Mert R. Sabuncu. Volumetric landmark detection with a multi-scale shift equivariant neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2020.

[51] M. Hamed Mozaffari, Noriko Yamane, and Won-Sook Lee. Deep learning for automatic tracking of tongue surface in real-time ultrasound videos, landmarks instead of contours. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, December 2020.

[52] Julia M. H. Noothout, Bob D. De Vos, Jelmer M. Wolterink, Elbrich M. Postma, Paul A. M. Smeets, Richard A. P. Takx, Tim Leiner, Max A. Viergever, and Ivana Isgum. Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE Transactions on Medical Imaging*, 39(12):4011–4022, December 2020.

[53] Jiahong Qian, Weizhi Luo, Ming Cheng, Yubo Tao, Jun Lin, and Hai Lin. CephaNN: A multi-head attention network for cephalometric landmark detection. *IEEE Access*, 8:112633–112641, 2020.

[54] Jiaxiang Ren, Heng Fan, Jie Yang, and Haibin Ling. Detection of trabecular landmarks for osteoporosis prescreening in dental panoramic radiographs. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine &amp Biology Society (EMBC)*. IEEE, July 2020.

[55] Imad Eddine I. Bekkouch, Tamerlan Aidinovich, Tomaz Vrtovec, Ramil Kuleev, and Bulat Ibragimov. Multi-agent shape models for hip landmark detection in MR scans. In Bennett A. Landman and Ivana Išgum, editors, *Medical Imaging 2021: Image Processing*. SPIE, February 2021.

[56] Kristina Belikova, Aleksandra Zailer, Svetlana V. Tekucheva, Sergey N. Ermoljev, and Dmitry V. Dylov. Deep learning for spatio-temporal localization of temporomandibular joint in ultrasound videos. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, December 2021.

[57] Bhargav J. Bhatkalkar, S. Vighnesh Nayak, Sathvik V. Shenoy, and R. Vijaya Arjunan. FundusPosNet: A deep learning driven heatmap regression model for the joint localization of optic disc and fovea centers in color fundus images. *IEEE Access*, 9:159071–159080, 2021.

[58] Xiaoyang Chen, Chunfeng Lian, Hannah H. Deng, Tianshu Kuang, Hung-Ying Lin, Deqiang Xiao, Jaime Gateno, Dinggang Shen, James J. Xia, and Pew-Thian Yap. Fast and accurate craniomaxillofacial landmark detection via 3d faster r-CNN. *IEEE Transactions on Medical Imaging*, 40(12):3867–3878, December 2021.

[59] Viacheslav V. Danilov, Kirill Yu. Klyshnikov, Olga M. Gerget, Igor P. Skirnevsky, Anton G. Kutikhin, Aleksandr A. Shilov, Vladimir I. Ganyukov, and Evgeny A. Ovcharenko. Aortography keypoint tracking for transcatheter aortic valve implantation based on multi-task learning. *Frontiers in Cardiovascular Medicine*, 8, July 2021.

[60] Junhyeok Kang, Kanghan Oh, and Il-Seok Oh. Accurate landmark localization for medical images using perturbations. *Applied Sciences*, 11(21):10277, November 2021.

[61] Hyuk Jin Kwon, Hyung Il Koo, Jaewoo Park, and Nam Ik Cho. Multistage probabilistic approach for the localization of cephalometric landmarks. *IEEE Access*, 9:21306–21314, 2021.

[62] Yankun Lang, Hannah H. Deng, Deqiang Xiao, Chunfeng Lian, Tianshu Kuang, Jaime Gateno, Pew-Thian Yap, and James J. Xia. DLLNet: An attention-based deep learning method for dental landmark localization on high-resolution 3d digital dental models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 478–487. Springer International Publishing, 2021.

[63] James McCouat, Irina Voiculescu, and Sion Glyn-Jones. Automatically diagnosing HIP conditions from x-rays using landmark detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2021.

[64] S. Palazzo, G. Bellitto, L. Prezzavento, F. Rundo, U. Bagci, D. Giordano, R. Leonardi, and C. Spampinato. Deep multi-stage model for automated landmarking of craniomaxillofacial CT scans. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, January 2021.

[65] Pavan Kumar Reddy, Aparna Kanakatte, Jayavardhana Gubbi, Murali Poduval, Avik Ghose, and Balamuralidhar Purushothaman. Anatomical landmark detection using deep appearance-context network. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp Biology Society (EMBC)*. IEEE, November 2021.

[66] L. C Tabata and C. N. Nyirenda. Faster r-CNN based cephalometric landmarks detection. In *2021 IEEE AFRICON*. IEEE, September 2021.

[67] Helena R. Torres, Pedro Morais, Anne Fritze, Bruno Oliveira, Fernando Veloso, Mario Rudiger, Jaime C. Fonseca, and Joao L. Vilaca. Anthropometric landmark detection in 3d head surfaces using a deep learning approach. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2643–2654, July 2021.

[68] Zhaohui Wang, Jun Shi, Xiaoyu Hao, Ke Wen, Xu Jin, and Hong An. Simultaneous right ventricle end-diastolic and end-systolic frame identification and landmark detection on echocardiography. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp Biology Society (EMBC)*. IEEE, November 2021.

[69] Qiang Zhang, Jixiang Guo, Tao He, Jie Yao, Wei Tang, and Zhang Yi. A novel landmark detection method for cephalometric measurement. In *2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE)*. IEEE, November 2021.

[70] Zhenwei Zhang, Shitong Mao, James Coyle, and Ervin Sejdić. Automatic annotation of cervical vertebrae in videofluoroscopy images via deep learning. *Medical Image Analysis*, 74:102218, December 2021.

[71] Samsuddin Ahmed, Kun Ho Lee, and Ho Yub Jung. Robust hippocampus localization from structured magnetic resonance imaging using similarity metric learning. *IEEE Access*, 10:7141–7152, 2022.

[72] Magnus Caspersen, Md. Sayed Tanveer, Asm Shihavuddin, M M Mahbubul Syeed, Md. Hasan Maruf, Ashraful Amin, and Faisal M. Uddin. Cascaded DNNs for detecting the position and orientation of left ventricle from 3d CT scans. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, June 2022.

[73] Dongfeng Du, Tao Ren, Chen Chen, Yiran Jiang, Guangying Song, Qingfeng Li, and Jianwei Niu. Anatomical landmarks annotation on 2d lateral cephalograms with channel attention. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, May 2022.

[74] Ali Pourramezan Fard, Joe Ferrantelli, Anne-Lise Dupuis, and Mohammad H. Mahoor. Sagittal cervical spine landmark point detection in x-ray using deep convolutional neural networks. *IEEE Access*, 10:59413–59427, 2022.

[75] Mohammad H. Jafari, Christina Luong, Michael Tsang, Ang Nan Gu, Nathan Van Woudenberg, Robert Rohling, Teresa Tsang, and Purang Abolmaesumi. U-LanD: Uncertainty-driven video landmark detection. *IEEE Transactions on Medical Imaging*, 41(4):793–804, April 2022.

[76] Cheng-Ho King, Yin-Lin Wang, Wei-Yang Lin, and Chia-Ling Tsai. Automatic cephalometric landmark detection on x-ray images using object detection. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, March 2022.

[77] Christoph Leitner, Robert Jarolim, Bernhard Englmair, Annika Kruse, Karen Andrea Lara Hernandez, Andreas Konrad, Eric Yung-Sheng Su, Jorg Schrottner, Luke A. Kelly, Glen A. Lichtwark, Markus Tilp, and Christian Baumgartner. A human-centered machine-learning approach for muscle-tendon junction tracking in ultrasound images. *IEEE Transactions on Biomedical Engineering*, 69(6):1920–1930, June 2022.

[78] Simon Schurer-Waldheim, Philipp Seebock, Hrvoje Bogunovic, Bianca S. Gerendas, and Ursula Schmidt-Erfurth. Robust fovea detection in retinal OCT imaging using deep learning. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3927–3937, August 2022.

[79] H Shankar, A Narayan, S Jain, D Singh, P Vyas, N Hegde, P Kar, A Lad, J Thang, J Atada, D Nguyen, PS Roopa, A Vasudeva, P Radhakrishnan, and S Devalla. Leveraging clinically relevant biometric constraints to supervise a deep learning model for the accurate caliper placement to obtain sonographic measurements of the fetal brain. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, March 2022.

[80] Zimeng Tan, Jianjiang Feng, Wangsheng Lu, Yin Yin, Guangming Yang, and Jie Zhou. Cerebrovascular landmark detection under anatomical variations. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, March 2022.

[81] Eyad Elyan, Pattaramon Vuttipittayamongkol, Pamela Johnston, Kyle Martin, Kyle McPherson, Carlos Francisco Moreno-García, Chrisina Jayne, and Md. Mostafa Kamal Sarker. Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. *Artificial Intelligence Surgery*, 2022.

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[83] Guangxing Han, Xuan Zhang, and Chongrong Li. Revisiting faster r-CNN: A deeper look at region proposal network. In *Neural Information Processing*, pages 14–24. Springer International Publishing, 2017.

[84] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2019.

[85] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation, 2016.

[86] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[87] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[88] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. 2019.

[89] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

[90] Omar Boudraa. Segmentation of 3d dental images using deep learning, 2022.

[91] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2016.

[92] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[93] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[94] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, feb 2015.

[95] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods, 2018.

[96] General information. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 1–2, 2015.

[97] Bastian Bier, Florian Goldmann, Jan-Nico Zaech, Javad Fotouhi, Rachel Hegeman, Robert Grupp, Mehran Armand, Greg Osgood, Nassir Navab, Andreas Maier, and Mathias Unberath. Learning to detect anatomical landmarks of the pelvis in x-rays from arbitrary views. *International Journal of Computer Assisted Radiology and Surgery*, 14(9):1463–1473, April 2019.

[98] Shandra Bipat, Saffire S. K. S Phoa, Otto M van Delden, Patrick M M Bossuyt, Dirk J Gouma, Johan S Lam??ris, and Jaap Stoker. Ultrasonography, computed tomography and magnetic resonance imaging for diagnosis and determining resectability of pancreatic adenocarcinoma: A meta-analysis. *Journal of Computer Assisted Tomography*, 29(4):438–445, jul 2005.

[99] Vajira Thambawita, Inga Strümke, Steven A. Hicks, Pål Halvorsen, Sravanthi Parasa, and Michael A. Riegler. Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. *Diagnostics*, 11(12):2183, November 2021.

[100] Martin J. Willemink, Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, and Matthew P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, apr 2020.

[101] Seong Ho Park and Kyunghwa Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, mar 2018.

[102] Jian Wang, Hengde Zhu, Shui-Hua Wang, and Yu-Dong Zhang. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26(1):351–380, nov 2020.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[104] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[105] Nick Hotz. What is crisp-dm? : Data science process alliance. Https://www.datascience-pm.com/crisp-dm-2/. Accessed: 2023-12-20.

[106] Özgün undefinediçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.

[107] Zimeng Tan, Jianjiang Feng, and Jie Zhou. Multi-task learning network for landmark detection in anatomical tree structures. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1975–1979, 2021.

[108] Trilinear interpolation - Wikipedia — en.wikipedia.org. https://en.wikipedia.org/wiki/Trilinear_interpolation#. [Accessed 08-07-2024].

[109] An Introduction to Biomedical Image Analysis with TensorFlow and DLTK — blog.tensorflow.org. https://blog.tensorflow.org/2018/07/an-introduction-to-biomedical-image-analysis-tensorflow-dltk.html. [Accessed 01-04-2024].

[110] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[111] Yuxin Wu and Kaiming He. Group normalization, 2018.

[112] 3D-UNet Medical Image Segmentation for TensorFlow | NVIDIA NGC — catalog.ngc.nvidia.com. https://catalog.ngc.nvidia.com/orgs/nvidia/resources/unet3d_medical_for_tensorflow. [Accessed 10-04-2024].

[113] François Chollet. *Deep Learning with Python*. Manning Publications, Shelter Island, NY, USA, 2018.

[114] Martin Kolarik, Radim Burget, and Kamil Riha. Comparing normalization methods for limited batch size segmentation neural networks. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, July 2020.

[115] Alex E. White, Peters T. Otlans, Dylan P. Horan, Daniel B. Calem, William D. Emper, Kevin B. Freedman, and Fotios P. Tjoumakaris. Radiologic measurements in the assessment of patellar instability: A systematic review and meta-analysis. *Orthopaedic Journal of Sports Medicine*, 9(5):232596712199317, May 2021.

[116] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, June 2018.

[117] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[118] Noura AlHinai. Chapter 1 - introduction to biomedical signal processing and artificial intelligence. In Walid Zgallai, editor, *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, Developments in Biomedical Engineering and Bioelectronics, pages 1–28. Academic Press, 2020.

[119] Gordana Sendić. Normal knee mri. https://www.kenhub.com/en/library/anatomy/normal-knee-mri, Nov 2023.

[120] Roberto M. Barbosa, Luís Serrador, Manuel Vieira da Silva, Carlos Sampaio Macedo, and Cristina P. Santos. Knee landmarks detection via deep learning for automatic imaging evaluation of trochlear dysplasia and patellar height. *European Radiology*, feb 2024.

[121] Neeraja R and L. Jani Anbarasi. A review on automatic cephalometric landmark identification using artificial intelligence techniques. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, nov 2021.

[122] What Is Human In The Loop | Google Cloud — cloud.google.com. https://cloud.google.com/discover/human-in-the-loop. [Accessed 11-06-2024].