UNIVERSITY OF LEEDS

Enhancing Total Hip Replacement Complications Diagnosis: A Deep Learning Approach with Clinical Knowledge Integration

Asma Alzaid

Submitted in accordance with the requirements for the degree of *Doctor of Philosophy*

The University of Leeds

School of Computing

September, 2023

Declaration of Authorship

I confirm that the work submitted is my own. The work which has formed part of jointly authored publications has been included. Chapters 3, 4, 5, and 6 of this thesis are derived from collaborative publications as the following:

- **Chapter 3:** "Automatic detection and classification of peri-prosthetic femur fracture", A. Alzaid, A. Wignall, S. Dogramadzi, H. Pandit, S. Q. Xie, 2022, International Journal of Computer Assisted Radiology and Surgery.
- **Chapter 4:** "Simultaneous Hip Implant Segmentation and Gruen Landmarks Detection", A. Alzaid, B. Lineham, S. Dogramadzi, H. Pandit, A. Frangi, S. Q. Xie, 2023, submitted to IEEE Journal of Biomedical Health Informatics (Accepted with minor revision).
- **Chapter 5:** "Medical Knowledge-based Learning for Peri-Prosthetic Fracture Diagnosis", A. Alzaid, S. Dogramadzi, H. Pandit, S. Q. Xie, 2023, Accepted in 29th IEEE International Conference on Mechatronics and Machine Vision in Practice.
- **Chapter 6:** "Reassembly of fractured object using fragment topology", Alzaid, A., and Dogramadzi, S., 2019, IET Conference Proceedings.

The co-authors played advisory roles by offering supervision and conducting reviews. All the original contributions outlined in these chapters are my own work.

I dedicated this thesis to...

The memory of my dear mother, whose passion for knowledge and unlimited support drove my pursuit of a PhD. Her greatest wish was to see me achieve this milestone, and though she is no longer with us, her spirit continues to inspire me every day.

My great dad,

My beloved husband,

My always supportive children Lina, Bader and Deema,

My lovely baby, Abdulrahaman, who came into the world during the end stages of this work.

Acknowledgments

I thank Allah, the almighty, for giving me this opportunity, the strength of endurance, patience and determination to complete my PhD thesis. The work presented in this thesis would not have been possible without the generous support and guidance of a number of people, to whom I am deeply grateful.

To begin with, I am very grateful to my first supervisor, Prof. Shane Xie, for providing me the opportunity to pursue my doctoral studies. His guidance, insightful advice, continuous support and motivation were instrumental in the successful completion of my work. I deeply appreciate the trust and freedom he granted to me to explore my own ideas and research paths. Also, my sincere appreciation goes to my co-supervisor, Prof. Hemant Pandit, for his insightful discussions, which deepened my clinical awareness, and for facilitating connections with specialists, crucial for generating valuable ground truth.

I would like to express my deepest gratitude to Prof. Sanja Dogramadzi, my co-supervisor, for her patience, trust, and support throughout my research journey. Her guidance and mentorship have been invaluable, and I am truly grateful for the confidence she has instilled in me. I appreciated her continuous discussions, inspiration, and readiness to provide advice and feedback whenever needed. Also, I would like to thank Prof. Alejandro Frangi for his invaluable advice and for introducing me to the numerous researchers within the community.

A huge thanks to my dearest father, whose wisdom, support and encouragement have been a constant source of strength and motivation. Likewise, I extend my appreciation and thanks to my beloved husband, Nasser, for his patience, unending support, and help at every step of this journey. I am also immensely thankful to my wonderful children, Lina, Bader and Deema for their love, understanding and patience during the long hours of work and research. A special thanks go to my lovely baby Abdulrahman, whose arrival filled my heart with joy and brought

my multitasking abilities to a new level. Also, I would like to thank my amazing siblings Eman, Abdulrahaman, Sumaya, Sara and Aisha, and my sister-in-law Bshayer for their unconditional love and support and for being there whenever needed.

A special thanks to my colleagues, Dr. Josh Lamb, Alice Wignall and Beth Linham, for their invaluable assistance during the data annotation phase and their guidance in comprehending various clinical-related intricacies. My dearest friends Nora Alkhamees and Nouf Bindiris thank you for your kind care, endless support, encouraged discussions and the unforgettable moments we shared in London, Bristol, Leeds and Colchester. I'm deeply grateful to all my friends in Bristol and Leeds for making the journey enjoyable and fun.

Last, I would like to acknowledge King Saud University for granting me a scholarship and financially supporting my research study.

Abstract

The increased rate of Total Hip Replacement (THR) for relieving hip pain and improving the quality of life has been accompanied by a rise in associated post-operative complications, which are evaluated and monitored mainly through clinical assessment of the X-ray images. The current clinical practice depends on the manual identification of important regions and the analysis of different features in arthroplasty X-ray images which can lead to subjectivity, prone to human error and delay diagnosis. Deep Learning (DL) based techniques showed outstanding outcomes across various image analysis tasks. However, the success of these networks is subjected to the availability of a very large, accurately annotated and well-balanced dataset - a constraint that is considered a main challenge for many medical image analysis tasks including THR.

This thesis focuses on automating the analysis of THR X-ray images to aid in the diagnosis and treatment planning of various THR complications. THR X-ray images including post-operation images and after Peri-Prosthetic Femur Fracture (PFF) images of a wide range of implants and various positioning and orientations, are collected to this end. Different Convolutional Neural Network (CNN) architectures are explored for PFF classification to observe how these networks perform in the presence of class imbalance and a limited number of data and with complex image patterns, either using full X-ray images or Region of Interest (ROI) images. This demonstrates that typical CNN-based methods succeeded in detecting PFF with DenseNet achieving an F1 score of 95%, while exhibiting low performance in the classification of PFF types, achieving an F1 score of 54% with GoogleNet, Resnet and DenseNet. This lower performance is attributed to the increased complexity of the task and the imbalanced distribution of the classes. To this end, the incorporation of THR medical knowledge with DL model is investigated.

The segmentation of the femoral implant component and the detection of important landmarks are formulated as simultaneous tasks within multi-task CNN that combines segmentation maps

of implant with the regression of shape parameters derived from the Statistical Shape Model (SSM). Compared to the state-of-the-art, this integrated approach improves the estimation of the implant shape by 6% dice score, making the segmentation realistic and allowing automatic detection of the important landmarks which can help in detecting many THR complications.

For PFF diagnosis, the incorporation of the clinical process of interpreting THR X-ray images with CNN is developed. For this purpose, the process of clinical interpretation of PFF X-ray images is defined and the method is designed accordingly. Four feature extraction components are trained to construct features from distinctive regions of the X-ray image that are defined automatically. The extracted features are fused to classify the X-ray image into a specific fracture type. The developed approach improved PFF diagnosis by approximately 8% AUC score compared to state-of-the-art methods, signifying notable clinical advancement.

Finally, the virtual pre-operative planning of bone fracture reduction surgery is explored which is important to reduce surgery time and minimize potential risks. The main obstacle toward the planning task is to define the matching between fragments. Therefore, 3D puzzle-solving method is formulated by introducing a new fragment representation and feature extraction method that improves the matching between fragments. The initial evaluation of the method demonstrates promising performance for the virtual reassembly of broken objects.

Contents

List of Figures

List of Tables

Abbreviation

THR	Total Hip Replacement
CADx	Computer-Aided Diagnosis
CADe	Computer-Aided Detection
CAP	Computer-Aided Planning
PFF	Peri-prosthetic Femur Fracture
DL	Deep Learning
CNN	Convolutional Neural Network
SSM	Statistical Shape Model
ReLU	Rectified Linear Unit
FC	Fully connected layers
AC	Active Contours
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CE	Canny Edge
ROI	Region of Interest
CAM	Class Activation Map
RPN	Region Proposal Network
FPN	Feature Pyramid Network
FCN	Fully Convolutional Networks

VOTT Microsoft Visual Object Tagging Tool

SGD Stochastic Gradient Descent

IOU Intersection Over Union

ASM Active Shape Model

VAE Variational Autoencoder

AP Antero-Posterior

GPA Generalized Procrustes Analysis

PCA Principle Component Analysis

MSE Mean Squared Error

NMSE Normalized Root Mean Square Error

CED Cumulative Error Distribution

RANSAC Random Sample Consensus

Chapter 1

Introduction

In 1991 it was suggested that Total Hip Replacement (THR) might be the operation of the century that can provide excellent pain relief and improve the quality of life for patients with hip pain or severe arthritis (Learmonth et al. 2007). With a growing elderly population, the rates of THRs are increasing, with approximately 90,000 procedures per year in the UK (United Kingdom National Joint Registry 2020), accompanied by an unavoidable rise in associated post-operative complications such as implant loosening, infection, dislocation or bone fracture. The main method to monitor recovery, assess THR complications and plan for further treatment relies on a clinical assessment of the patient, prior operation notes on the joint implant and a surgical approach taken and evaluating the hip X-ray images.

Medical image analysis has advanced significantly in recent years, transforming the way of interpreting, processing and utilizing medical images. These advancements can be categorized into three major areas: Computer Aided Detection (CADe), Computer Aided Diagnosis (CADx) and Computer Aided Planning (CAP). Each of these areas targets specific problems and is essential for improving treatment outcomes. CADe focuses on the identification and localization of specific abnormalities within an image, such as defining if there is a fracture in the X-ray image or not. CADx includes not only the detection of abnormalities but also the characterization and classification of different types of abnormality which provides a comprehensive evaluation of the patient's condition, such as defining the specific type of the fracture from the X-ray image. CAP formulates treatment strategies and surgical plans using imaging data which can improve treatment delivery and minimize risks.

The current assessment of THR X-ray images depends on the expertise of clinical specialists to analyse intricate details in the image which can lead to outcome variability, human errors, and potential delay in diagnosis. In addition, the shortage of experienced clinical specialists and the rising demand for THR treatments further complicate the issue. For instance, Marshall et al. 2017 indicated that 90% of Peri-prosthetic Femur Fracture (PFF) radiograph reports in the emergency department do not include all relevant radiographic features, which can delay diagnosis and treatment. Proposing a THR CADx tool aims to minimize these challenges by providing an automated approach to image analysis, enhancing accuracy and efficiency in diagnosing THR-related complications. In addition, it can reduce the potential for human error and subjectivity, contribute to the earlier detection of diseases and ultimately improve patient outcomes. While there is limited empirical research on detecting THR complications compared to the huge development in other medical domains, the existing studies often face challenges in terms of reliability and generality due to the dataset problems including the imbalanced class distribution, limited size dataset and the quality of used images (Patel et al. 2021; Borjali, A. F. Chen, Muratoglu, et al. 2019; Khosravi et al. 2022; Rouzrokh, Ramazanian, et al. 2021). The scope of these studies is restricted to the detection of only one type of condition and does not provide insights into the severity and classification of that condition. This further limits the existing research.

Another critical stage in orthopaedic surgery is pre-operative planning, particularly in fracture reduction operations. It includes a detailed analysis and planning before the surgery, aiming to achieve optimal outcomes and minimize complications. The development of 3D object modelling and processing techniques has enhanced computer-assisted fracture preoperative planning. However, finding the matching between broken bone fragments is still challenging due to the following reasons: the matchable properties between fragments are located surrounding the fractured area only as opposed to a full or partial matching problem where there are notable overlapping patterns. Also, missing parts of fragments will significantly complicate the problem. Developing a 3D fragment reassembly tool can enhance the fracture reduction pre-operative planning which ultimately improves the treatment outcomes.

Therefore, this thesis investigates the effectiveness of CADx systems for the diagnosis of THR complications such as PFF. Additionally, the exploration of how the integration of medical knowledge into Deep Learning (DL) impacts the enhancement of accuracy and consistency in

THR image analysis and to what extent this integration can overcome dataset limitations is undertaken. Moreover, an investigation is conducted on how computer-based methods can be employed to assist in the pre-operative planning of fracture reduction surgery.

1.1 Research Aim and Objectives

The primary aim of this thesis is to formulate an automatic and robust framework for THR X-ray image diagnosis, to enable the analysis of intricate information and other features of interest and assist in treatment planning. Guided by the above aim, the following objectives are defined :

1. Assess the typical Convolutional Neural Networks (CNNs) architecture performance on THR X-ray image analysis task with dataset limitation and challenges.
2. Identify the important landmarks in the implant component and automate the detection of these landmarks and the segmentation of the implant.
3. Develop a framework for the diagnosis PFF from X-ray images, addressing current data limitations and associated challenges.
4. Design a framework for fracture reassembly to assist in the pre-operation planning for fracture reduction

1.2 Outlines and Contributions

The work presented in this thesis develops diagnostic methods for THR-related complications from X-ray images. Developed methods are focused on the PFF diagnosis due to the absence of an automated method in this domain and the complexity of the task compared to the detection task only. DL-based methods are designed for automated detection and diagnosis of PFF and for defining important landmarks in THR X-ray images. In particular, this thesis creates new methods that integrate medical knowledge and DL model that improve diagnostic and image analysis even in the presence of dataset limitations and THR X-ray image challenges.

This thesis consists of this introductory chapter followed by a chapter reviewing the background and the state-of-the-art in DL classification and segmentation tasks within the field of medical image analysis. Four research chapters reporting all the results of this thesis are followed by a

final chapter which presents the conclusion and the future directions.

- **Chapter 2** is a background and literature review that provides a medical background of THR and explains its associated complications and the variations among THR patients as well as a state-of-the-art image classification and segmentation, particularly for THR. Finally, the limitations and challenges of these techniques, including those in the dataset, are defined.

- **Chapter 3** reports a method developed using CNN to detect, localize and identify the PFF classification from X-ray images. A large dataset of post-operative THR X-ray images and after the PFF X-ray images have been collected and annotated. Also, this chapter discusses the existing work on bone fracture detection and diagnostic methods. The results of different CNN architectures for the three tasks, detection, localization and diagnosis based on clinical standards were compared. These architectures have been further evaluated for situations that involve limited data availability and imbalanced class distribution. The challenges and the variation of PFF X-rays compared to the other bone fractures are also demonstrated in this chapter.

- **Chapter 4** provides results of an integrated approach that enables the segmentation of implant components and the detection of the landmarks of interest. In addition, the landmarks of the Gruen zone are defined and the shape of the implant is represented correspondingly. This can aid the identification of many THR complications that are computed according to these landmarks. In more detail, this chapter discusses the existing methods in implant segmentation and the existing approaches in integrating medical knowledge with DL models in various medical domains. The anatomical knowledge and the landmarks of interest are explained as well as the developed method for implant segmentation and Gruen landmarks detection. This chapter demonstrated that combining segmentation with landmarks detection improves the results of landmarks detection and segmentation of the implant even with the dataset size limitation.

- **Chapter 5** reports results of a novel medical knowledge DL network for PFF diagnosis that enables the extraction and fusion of features from the most discriminated regions that a clinical specialist focused on when interpreting PFF X-ray images. This approach improves the PFF diagnosis performance compared with the available CNNs. In more detail, this chapter discusses the existing work on fracture detection and diagnosis explains

and defines the clinical process of reading the PFF X-ray images. The developed method combines multiple features from the X-ray images and simulates the clinical diagnosis. The reported results demonstrated improvement in PFF diagnosis compared with the state-of-the-art methods.

- **Chapter 6** develops a method for reassembly of the broken 3D objects for the purpose of aiding the pre-operative planning for fracture reduction operation. The main contributions of this chapter are the proposal of a reassembly method that combines the intact facet features with the fractured facet boundary to minimize the potential matching between fragments, the introduction of a new representation of the fragment that represents the fragment features and simplifies the search for possible matching, the proposed new feature that improves the segmentation and classification of the fragment facets significantly. In more detail, the existing approaches for reassembling fractured objects in different domains are presented, such as bone fracture reduction and broken artefacts. Then, the steps of the developed reassembly method including the segmentation of the fragment surface into intact and fractured facets, the generation of fragment topology and descriptors and the matching process are explained. Also the evaluation process and discussion of the results of each step are presented.
- **Chapter 7** concludes this thesis by summarising and discussing the main findings and outlining the plan for future work.

1.3 Publications

The work conducted during this PhD project has led to the following publications:

1. A. Alzaid, A. Wignall, S. Dogramadzi, H. Pandit, S. Q. Xie, "Automatic detection and classification of peri-prosthetic femur fracture," International Journal of Computer Assisted Radiology and Surgery, vol. 17, pp. 649–660, 2022.
2. Alzaid, A., Lineham, B., Dogramadzi, S., Pandit, H., Frangi, A.F. and Xie, S.Q., 2023. Simultaneous Hip Implant Segmentation and Gruen Landmarks Detection. IEEE Journal of Biomedical and Health Informatics.
3. A. Alzaid, S. Dogramadzi, H. Pandit, S. Q. Xie, Medical Knowledge-based Learning for Peri-Prosthetic Fracture Diagnosis. Accepted at 29th IEEE International Conference on

Mechatronics and Machine Vision in Practice. Achieved John Billigsley best conference paper award on Computer Vision

4. Alzaid, A., and Dogramadzi, S. (2019). Reassembly of fractured object using fragment topology. IET Conference Proceedings, 2019, p. 18 (98 -105), DOI: 10.1049/cp.2019.0256.

Chapter 2

Background

2.1 Total Hip Replacement

Femur is the largest bone in the body and is partitioned into three main parts: proximal, shaft and distal. Figure 2.1 shows the normal femur anatomy. The proximal femur is composed of a head, neck, and greater and lesser trochanters. The head is articulated with the pelvic bone acetabulum to form the hip joint. All the body load during daily activities is transferred through the acetabula onto the femoral heads. Both greater and lesser trochanter are attachment sites for muscles that allow movement of the hip joint. The femur shaft has a cylindrical shape which differs significantly between individuals. The shaft anterior is smooth and lacks distinctive features. However, the opposite side is rougher and consists of Linea aspera that is raised linearly along the long access of the femur and split to form the medial and lateral supracondylar lines. The Distal femur is the widest part of the bone. It is characterized by the lateral and medial condyles that connect with both the tibia and patella to compose the knee joint.

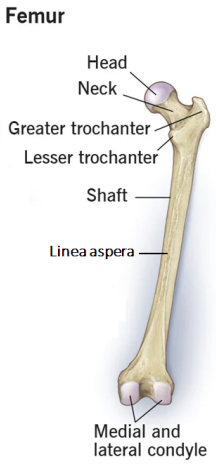


Figure 2.1: Femur bone anatomy.

THR is a surgical operation that replaces the hip joint with a prosthetic implant. This includes replacing the acetabulum and femoral head. Generally, THR is conducted to mitigate arthritis pain or to treat extreme joint damage after a hip fracture. This will improve the quality of life for patients. Figure 2.2 shows the component of THR. The damaged femur head is removed

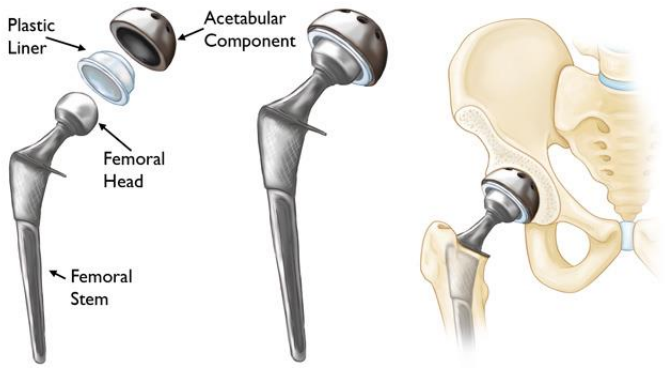


Figure 2.2: The component of THR (Foran and Fischer 2020)

and the femoral stem is fitted in the centre of the femur. The THR surgery can be carried out on young people who suffer from severe arthritis pain or complex hip fracture complications, but it is mainly performed among people aged 60 years and above (OECD 2019). Over the past decade, the number of THR has increased rapidly, around 30% between 2007 and 2017 (OECD 2019). According to the United Kingdom National Joint Registry 2020, approximately 90,000 THR operations are performed per year in the UK.

Although THR results in excellent pain relief for patients and improves their quality of life, it is associated with many risks such as infections, bone fractures, implant loosening and implant subsidence (United Kingdom National Joint Registry 2020). Routine follow-up visits and early detection could minimise the effect of these complications. The main assessment method for postoperative THR is through medical images. While MRI, CT, and sonography are commonly used for joint imaging, the postoperative radiograph remains useful in evaluating hip arthroplasty due to its effective cost, readily available, free from metal artefacts and enables longitudinal comparison (Vanrusselt et al. 2015).

The hip prosthesis can significantly vary among patients due to many factors. For instance, there are two surgical options for femoral stems: cemented and cementless stems. On cemented stems, the prosthesis is fitted in position securely using special bone cement, while on cementless stems, the implant is fitted into the proximal femur and the surrounding bone grows into the prosthesis to form a solid attachment and on the upper part of the stem a ball, metal or ceramic, is placed which replaces the damaged femur head. Also, the acetabulum bone is removed and replaced with a metal socket and a spacer is inserted between the ball and the socket. Other hip prosthesis variation factors include design variations, material choices, manufacturers, patient-specific considerations and geographic and demographic factors. These variations increase the

complexity of THR X-ray image analysis.

2.2 Medical Image Analysis

Since the invention of X-ray images in 1895, medical imaging has become essential in diagnosing lots of diseases. It provides information about the patient's medical condition and indications of the causes of the symptoms and diseases. The development of medical image technology led to an increase in the number of producing, which radiologists must interpret. The effectiveness of image interpretation by skilled medical professionals is impacted by the limited number of available experts, as well as the challenges posed by their exhaustion and the reliance on approximate estimation procedures. Consequently, the importance of medical image analysis tools extends across various areas of medicine including the detection and diagnosis of disease. Medical image analysis has overcome traditional diagnostic limits by collecting quantitative data and visualising anatomical characteristics that might be undetectable to the human eye.

Two key tasks of medical image analysis will be explored: image classification and image segmentation. These methods are essential for CADe and CADx tools supporting the translation of unprocessed image data into useful information that assists clinical specialists in making decisions, treatment planning and research advancement. Image classification includes the categorization of medical images into predefined classes, allowing the identification of specific conditions, abnormalities, or characteristics. This technique is significant in medical diagnosis by aiding clinicians in making accurate and timely decisions. Image segmentation involves the division of an image into meaningful regions or structures that facilitate the localization of anatomical structures or regions of interest within an image.

2.2.1 Image Classification

Image classification involves training algorithms to classify images into predefined classes based on the visual content. These algorithms learn to recognise different patterns and features that distinguish one class from another using machine learning techniques which enable the algorithm to identify the labels of new images. Let I denote an image consisting of pixels and let $l_1, l_2, ..., l_k$ denote class labels. For each pixel x, a feature vector V including values $f(X_i)$ is extracted using:

$$V = (f(X_0), f(X_1), ...f(X_n)) \tag{2.1}$$

Where n is the number of pixels. A label l_i is assigned to the image based on V. The process of image classification includes the partitioning of dataset into training, validation and testing sets. The training set is used for model training, a validation set is used for parameter tuning, and a test set for evaluating the model's performance.

Image classification methods can be categorized into two groups: classical methods and DL-based methods. The classical methods involve standard machine learning techniques such as Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN) which commonly require hand-engineered features and class labels to classify images. This can limit their ability to recognise complicated patterns in medical images. The emergence of DL has improved image classification and CADx significantly by learning hierarchical features from raw data automatically. A gradual transmission from typical methods of computer vision to DL based techniques has recently been observed. Beginning in 2012, DL showed great success in several computer vision tasks involving the classification, detection and matching of images. After that, treating more complicated tasks, such as medical image analysis, was only a matter of time.

The foundation of DL for image analysis is the CNN. There are many CNN architectures applied for medical image classification. The widely utilized architectures are AlexNet (Krizhevsky et al. 2012, VGG (Simonyan and Zisserman 2015), GoogleNet (Szegedy et al. 2015) , ResNet (K. He et al. 2016) and DenseNet (G. Huang et al. 2017). The following paragraphs provide a general overview of the CNN architecture.

A CNN takes an image as input and assigns importance (learnable weights and biases) to different elements/objects in the image enabling the distinction between images. Classic CNNs consist of three main components: convolutional layers, pooling layers (or subsampling layers), and fully-connected layers (FC) that are stacked to form the architecture. Advanced architecture could involve more layers such as batch normalisation and dropout layers. The input and output of each layer are sets of arrays called feature maps.

The convolutional layer is the key element of CNN which captures local patterns and complex features by sliding filters (kernels) across the input image. The parameters of each layer involve kernel size, stride size and padding that are applied to the side of input feature maps. These parameters are responsible for determining the weight of the learnable convolutional kernels.

The kernel is applied to small localised regions of the input image or the feature map of the

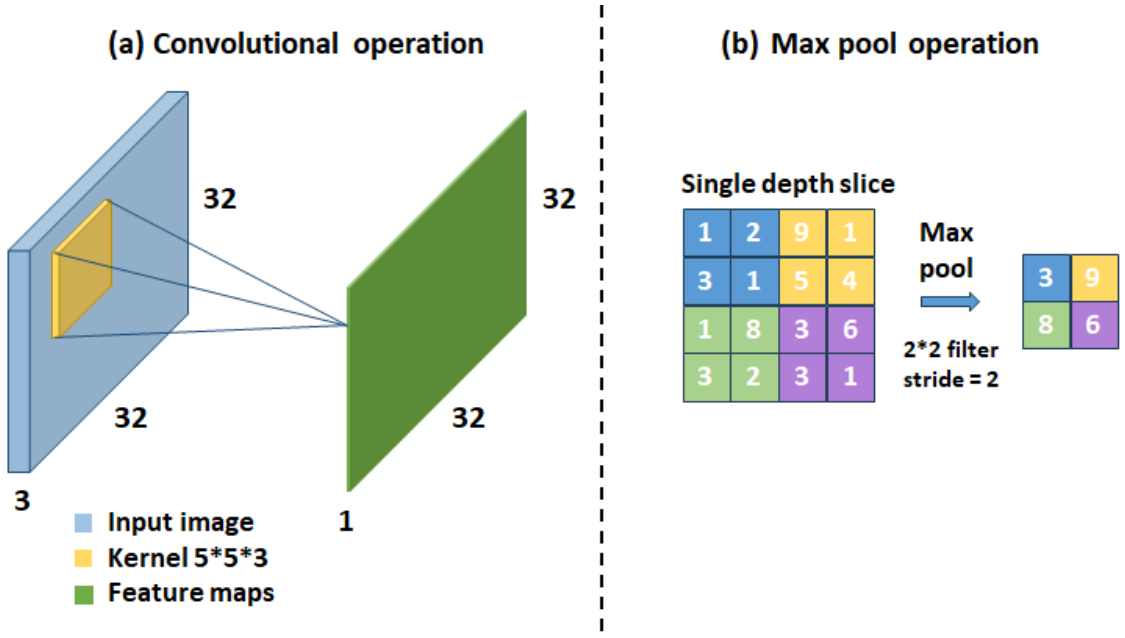


Figure 2.3: (a) an example of the convolutional operation. (b) an example of pooling operation using max pool.

previous layer and is moved in both horizontal and vertical directions. The moving of the kernel from the top left or bottom right corners is termed the stride. When a large stride is used, the lesser filter will be applied, which produces a smaller output size, and vice versa. Figure 2.3 (a) presents the convolutional operation. The output values of the convolutional operation pass through an appropriate activation function such as Rectified Linear Unit (ReLU) and sigmoid, to capture complex patterns and relationships within data.

The pooling layers followed the convolutional layers. It performs a down-sampling operation resulting in an equal number of feature maps that are given as input to the next layer. Also, pooling makes the representation invariant to small translations and rotations. The most common pooling function are Max pooling, which returns the maximum value of the receptive field, and Average pooling, which return the average of a local neighbourhood. This layer is also identified by kernel size, stride size and padding. Figure 2.3 (b) shows the pooling layer function.

Fully connected layers (FC) are added on the top of convolutional and pooling layers. It takes as input the output of all units of the previous layer. The first FC layer usually transforms the output of the last convolutional or pooling layer to a vector of predefined length. Then, this vector can either be an input to a classifier or be considered as a feature vector to solve a different task.

Dropout layer is a regularization technique used to prevent overfitting. In the training phase, dropout randomly deactivates some neurons which helps the network to learn more diverse and representative features.

Batch normalization layer is used to stabilize and accelerate the training process by normalizing the outputs of a layer through re-centring and re-scaling. This helps to avoid the vanishing or inflating gradient problem, making learning more steady and efficient. Another advantage of a batch normalization layer is the use of a larger learning rate without a vanishing gradient problem. In addition, the impact of regularizing enhances the network generalization and minimizes overfitting.

CNNs have made significant contributions to medical image classification including the detection and diagnosis of abnormalities. These have been investigated and effectively applied for a variety of medical image interpretation applications like detecting and diagnosing lung cancer and respiratory diseases such as pneumonia, tuberculosis and COVID-19 (Meedeniya et al. 2022, Çallı et al. 2021), heart diseases (Vesal et al. 2020,), Breast tumours (Yan et al. 2020), Brain diseases (Dhiman et al. 2022). The growth of these medical applications is facilitated by the availability of dataset such as chestX-ray14 (X. Wang et al. 2017), COVID-19 Radiography Dataset (Chowdhury et al. 2020) and MIT-BIH Arrhythmia database (Goldberger et al. 2000). On the other hand, numerous medical image domains have limited investigations compared to those applications such as bone diseases and dental X-ray images. This is mainly due to the limitation of the available dataset.

The techniques introduced for CNN-based medical image classification are varied and can be categorized based on their architectural approach. These include the utilization of pre-existing CNN models, leveraging transfer learning (Maghdid et al. 2021), employing ensemble models (Q. Liu et al. 2020) and employing a hybrid approach (Dhiman et al. 2022) etc.

Despite the huge improvement in medical image classification with the emergence of CNN approach, many challenges still exist in this field. Section 2.2.3 discusses these challenges.

2.2.2 Image Segmentation

Image segmentation is the process of partitioning an image into distinguished, meaningful segments that help to identify important objects or structures and extract useful information from an image. Many hand-engineered feature-based algorithms are utilized for image segmentation.

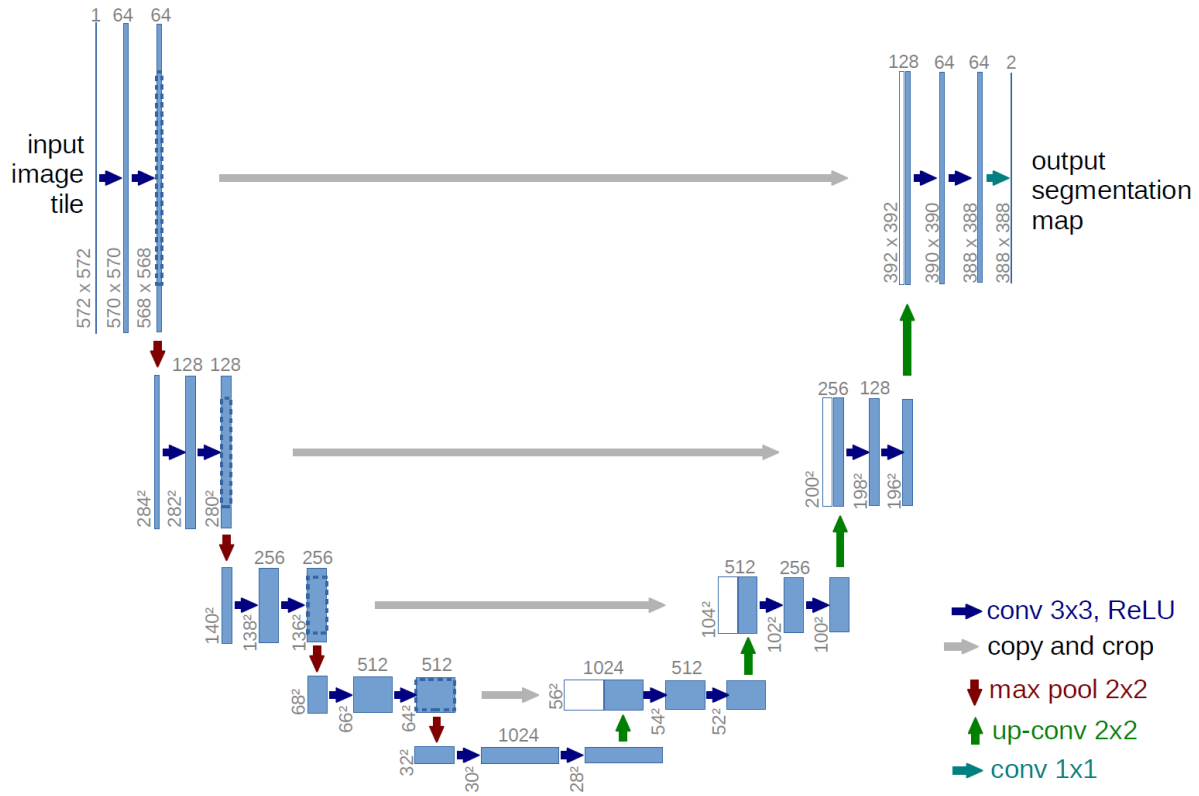


Figure 2.4: U-Net architecture (Ronneberger et al. 2015)

These techniques vary from simple thresholding or region-based methods to more advanced methods such as Atlas-based methods and Statistical Shape Model (SSM) (Gan et al. 2021). However, the development of DL methods, particularly CNNs, has influenced the advancement of image segmentation significantly even with complex data such as medical images (Tajbakhsh et al. 2020). It provided state-of-the-art results in many image segmentation tasks.

The typical form of using CNNs for image segmentation is the same as the classification of each pixel in the image. This is known as semantic segmentation where each pixel is associated with a label or category. U-Net (Ronneberger et al. 2015) is one of the common segmentation architectures due to its effectiveness in medical image segmentation. U-Net features allow the merging of low-level and high-level features while keeping both detailed information and contextual knowledge.

U-Net Architecture U-Net is a CNN-based architecture that is designed specifically for image segmentation tasks. U-Net name is derived from its structure which is a unique U-shape structure. Figure 2.4 illustrates the U-Net architecture. It consisted of two main parts encoder path (left side) and decoder path (right side). The encoder side is similar to the convolutional network architecture, consisting of two convolutions each followed by ReLU and 2×2 max

pooling with stride 2 for downsampling. For each downsampling step, the number of feature channels is doubled. Each step on the decoder side involves an up-sampling of the feature map followed by 2×2 convolution which splits the number of feature channels by two, a connection with the corresponding cropped feature map from the encoder side, and two 3×3 convolutions each followed by ReLU. The final layer consists of 1×1 convolution which maps each 64 features vector to a predefined number of classes.

CNNs have made significant advancements to medical image segmentation by improving accuracy, minimizing manual intervention and applicability to various modalities. These have been investigated and effectively adapted for several medical image segmentation domains like brain tumor segmentation (Naser and Deen 2020), breast tumors (Ranjbarzadeh et al. 2022), organs segmentation such as livers (Kavur et al. 2020) and pancreas (Zheyuan Zhang et al. 2023).

2.2.3 Challenges and limitations of DL in medical images

Medical image analysis has been transformed by DL, yet it still has a number of challenges and limitations. The following points discuss and summarise some of these issues:

1. **Data limitation:** CNN models require large and well-annotated datasets for successful training. In the medical image domain, collecting and labelling data can be expensive and time-consuming. Many approaches have been used to automate the labelling process such as the use of natural language processing for label extraction from images. However, the use of expert radiologist labels is recommended for accurate assessment (Çallı et al. 2021). Using a limited dataset can lead to overfitting and reduced generalization. Several techniques can be used to overcome dataset limitations such as augmentation and transfer learning, however, a well-trained model still requires a large and well-labelled dataset (Kwong and Mazaheri 2021). For instance, some augmentation methods such as rotations and flips maintain the information of the data.

2. **Class imbalance:** is a common property of medical datasets, where some diseases or conditions are more common than others. This imbalance can lead to learning bias, causing predictions to favour the overrepresented class. This can also impact the metrics used for assessing the model like accuracy. Different methods can be used to overcome the class imbalance. These methods can be applied on the data level, on the model level or on both (J. M. Johnson and Khoshgoftaar 2019), however, the effectiveness of these

methods depends on many factors including dataset size, level of imbalance, quality of features and the applied method.

3. **Computational cost:** DL training and running can be highly computational, posing memory constraints, particularly when dealing with large feature space training data such as high dimensional or high-resolution images. Implementing DL in clinical settings with limited resources can be challenging.

4. **Interpretability:** In the medical image domain, interpretability and explainability are essential to visually validating the proposed tool. Many studies use methods like CAM or saliency maps (Bolei Zhou et al. 2016) to show the area of interest. These heatmaps might help in showing conditions that appear as localized patterns, however, a comprehensive evaluation of their accuracy is commonly lacking. On the other hand, many conditions are difficult to visualize through these methods due to difficulties in explaining these conditions.

In spite of these challenges, DL approaches continue to improve the medical image analysis field. Future research aims to overcome and minimize these limitations.

2.3 Medical Image Analysis in THR

Numerous computer-based methods have been introduced in the literature to aid the detection and diagnosis of THR complications. This section provides an overview of these methods.

The essential step for analysing THR X-ray images is the segmentation of the implant. Many image segmentation techniques have been developed to segment the implant from the X-ray image. Oprea and Vertan 2007 evaluated several classical adaptive region segmentation techniques, using either adaptive histogram thresholding or an extended feature space (Fuzzy C-mean clustering). In the fuzzy C-mean clustering they assumed that the image consisted of a combination of three classes (prosthesis, bone and soft tissue). However, the grey-level-based methods are unstable due to the high impact of radiometric distortion in X-ray images. Similarly, L. Florea and Vertan 2009 attempted to find the optimal thresholding value by applying the Expectation Maximization algorithm to segment the X-ray images into three classes. Then, they applied a Canny Edge algorithm (CE) to detect and separate the prosthesis from the bone. Downing et al. 1997 segmented the hip prosthesis relying on the fact that a prosthesis is a high-contrast object.

Thus, they first applied the CE method to detect edges in the X-ray image. Then, edges were thresholded and connected to produce close regions that were classified by comparing simple features such as brightness, perimeter and area of the prosthesis. Femur shaft access was detected using the distal end of the prosthesis and the prosthesis head was detected by extracting Circular Hough Transform features. Then a template was matched to extract the femur shape. Finally, the soft tissue, bone, cement and prosthesis are segmented and classified based on the intensity profile of the X-ray image. Recently, the Active Contours (AC) method has been used to segment implant in the hip (Al-Zadjali 2017, Jia et al. 2011). The seed point for the AC method was initialized based on the detected points in the prosthesis cup. Al-Zadjali 2017 utilized the Fast randomized circle detection (Jia et al. 2011) to detect the centre and radius of the prosthesis cup, while Stark 2018 applied Hough transforms method to detect circles in the image to localize the implant in the shoulder. Then a region-growing method was used to segment the implant. The hand-crafted-based segmentation can provide good results when the implant components are well-defined in the X-ray images. With the huge improvement of the DL method, CNN-based techniques have been applied to segment implant in the X-ray image such as Unet (Patel et al. 2021).

On the other hand, various methods were proposed to identify complications or risks related to THR. These methods were based on image processing methods, DL methods or both. Table 2.1 presents an overview of these methods. For implant subsidence detection, Barker and Donnelly 2003 developed a semi-automatic detection method that consists of multiple image processing steps. First, two landmarks are marked manually in the distal centralizer and in the cancellous bone proximal. Circular Hough transformation and Sobel edge operator were used to localize the centre and radius of the femoral head component. The projection of the vertical and horizontal edge of the implant was computed using the Random transform and the amount of migration was computed using the derived intersection points. Their method was applied to one type of implant only and depended on the manual positioning of landmarks. Al-Zadjali 2017 also proposed an image processing-based method for the detection of femoral component subsidence and loosening by defining the region of interest based on curvature analysis. Their method started using the segmented implant which was described earlier in this section. Then, the extraction of boundary pixels was conducted using image pre-processing steps such as generating a binary image and edge map from the segmentation and tracing the edge by the Edge link algorithm. After that, a curvature analysis was utilized to extract the key points on the implant,

which were the points with significantly varying curvature along the boundary. The loosening and subsidence were detected using pixel gradients at the extracted points. Image processing-based methods are limited to simple anatomical structures and their effectiveness depends on image quality. Rouzrokh, Wyles, Kurian, et al. 2022 enhanced these limitations by using CNN based method for implant subsidence detection. First, a dynamic U-Net model was trained to segment femur, implant and magnification markers followed by an image processing method to measure subsidence. The image processing steps computed the standard femur stem distance on each radiograph to detect the subsidence. It received the output segmentation masks and computed the coordinates of the tip point of the implant (stem point) and the coordinates for the most superior point on the greater trochanter (femur point), and calculated the standardized distance between them based on the angle of the femoral axis and the distance between the magnification markers. The final step involved calculating stem subsidence by subtracting the measured standard femur-stem distance between the two images. Their method considered patients with polished tapered cemented femoral stems only.

For implant dislocation risk detection, Rouzrokh, Ramazanian, et al. 2021 aimed to detect patients with a high risk of implant dislocation by analyzing postoperative THR X-ray images with dislocations over 5 years of follow-up. They trained YOLO-V3 to crop the images by centring on the femoral head and then ResNet18 was trained to classify images with dislocated risk. However, they obtained a very low positive predictive value of 3.3. Khosravi et al. 2022 enhanced the overall detection performance by combining image features with clinical information. A hybrid network, consisting of EfficientNet-B4 and Swin-B transformer models, was utilized to extract image features. These features were combined with clinical characteristics such as demographics, comorbidities, and surgical details. The combined features were used to train an XGBoost model to identify the risk of dislocation. The results demonstrated that incorporating clinical information improved the overall detection performance. In addition, Rouzrokh, Wyles, Philbrick, et al. 2021 proposed a method to detect acetabular component inclination from THR X-ray images using both Antero-Posterior (AP) and Lateral radiographs. Two U-nets were trained, one for AP and one for lateral, to segment the acetabular component and the ischial tuberosities. Multiple image processing steps were deployed to measure the acetabular component angles using the resulting segmentation mask. They demonstrated limitations in patient positioning, including rotations over 45 degrees, cropped images lacking key components, and hidden ischial tuberosities due to soft tissue overlap or unusual hardware presence. Moreover,

the lack of control for pelvis positioning before measurements impacted their method's reliability and applicability in real-world scenarios with diverse radiographic conditions.

Additionally, several methods were proposed to define the type of implant design using post-operative THR X-ray images. Many of these methods trained typical CNN-based models such as DenseNet and Resnet to classify the THR X-ray image to specific kind of implant design (Karnuta et al. 2021; Borjali, A. F. Chen, Bedair, et al. 2021; Gong et al. 2022). On the other hand, other methods used the cropped region of the implant as input to train the classification model. Kang et al. 2020 first trained YOLOv3 to detect the stem component region and then developed a simple CNN architecture to identify the specific implant type. Patel et al. 2021 trained U-net to segment the implant mask and used the mask to crop implant region. Their CNN architecture consisted of two components that were jointly connected for final classification. The first component utilized the full X-ray image, while the other component focused on the segmented region of the implant. Each of the previous methods employed a different set of implant design categories for instance Karnuta et al. 2021 defined 18 implant designs, while Kang et al. 2020 defined 29 types.

Furthermore, many existing methods focused on detecting implant loosening in THR X-ray images through training typical CNN models (Borjali, A. F. Chen, Muratoglu, et al. 2019). Shah et al. 2020 studied the impact of other input variables on the performance of typical CNN. They trained DenseNet on THR X-ray images that were labelled as 'loose' or 'well-fixed' based on the operation notes. In addition, historical, demographic, and comorbidity information were added to create a final model. Their results showed that when combining historical data and fine-tuning the pre-trained DenseNet, the accuracy was improved significantly. On the other hand, T. Rahman et al. 2022 leveraged the strengths of different CNN models and classifiers to improve the overall prediction performance of the detection of Aseptic loosening in THR X-ray images by developing a stacking ensemble method. First, YOLOv5 was trained to localize and crop implant regions. DenseNet201, Resnet50, and Resnet18 were used as base learners for the stacking ensemble and the prediction outputs were used as input to Random Forest, which is a meta-learner. They collected 206 THR X-ray images and reported a detection accuracy of 96%.

On the other hand, several studies introduced the detection of multiple complications by labelling it as implant failure. Loppini et al. 2022 trained DenseNet to identify implant failure from THR X-ray images including loosening, malpositioning of the implant, polyethylene wearing, or

peri-prosthetic infection. Muscato et al. 2023 developed a method for detecting implant failure from postoperative THR X-ray images including loosening, bearing surface wear and osteolysis, malpositioning and dislocation. Their method started with pre-processing steps to enhance the X-ray image and segment the image into the acetabulum region and the stem region. These steps included applying gamma power transformation, sigmoidal function, contrast-limited adaptive histogram equalization (CLAHE), and low pass filter with 2D Gaussian smoothing kernel. In addition, the CE algorithm was used to identify the presence of the implant and to segment the acetabulum region (upper third) and the stem region. Each of the three images; full, acetabulum and stem were utilized as input to three Densenet models to extract image features. The obtained features were reduced using PCA method and concatenated and fed to the SVM algorithm for classification.

Reference	**Purpose**	**Method**	**Dataset**
Barker and Donnelly 2003	subsidence complication	image processing based method	-
Al-Zadjali 2017	subsidence and loosening complications	image processing based method	-
Rouzrokh, Ramazanian, et al. 2021	Dislocation Risk	YOLO-V3 : crop image-centering on femoral head. Resnet18: classification	92,584 radiograph (1490 dislocated / 91,094 non-dislocated)
Khosravi et al. 2022	dislocation risk	Swin-B and EfficientNet-B4 for image feature extraction. XGBOOST classification	170,73 radiographs + demographic data (2% dislocated samples)
Borjali, A. F. Chen, Muratoglu, et al. 2019	implant design (9 designs)	DenseNet201	402 radiograph (130 Accolade II, 89 Corail, 31 M/L Taper, 31 Summit, 26 Anthology, 26 Versys, 24 S-ROM, 24 TSO, 21 THO)

Reference	**Purpose**	**Method**	**Dataset**
Rouzrokh, Wyles, Philbrick, et al. 2021	dislocation	dynamic Unet + Image processing	500 radiograph (Polished tapered cemented femoral stem only)
T. Rahman et al. 2022	Aseptic loosening	YOLO-v5: crop implant region. Stacking ensemble: classification	216 radiograph (112 loose / 94 controlled)
Muscato et al. 2023	implant failur: loosening, dislocation, bearing surface wear and osteolysis and malpositioning.	image processing: enhancing and segmentation. DenseNet102 : feature extraction. SVM: classification	1331 radiograph (869 failed / 462 non-failed)
Gong et al. 2022	implant design (4 types)	ResNet50	710 radiograph (350 Biomet Eco Bi-Metric, 210 Depuy Corail, 90 LINK Lubinus SP II, 60 Zimmer Versys FMT)
Borjali, A. F. Chen, Bedair, et al. 2021	Mechanical loosening	DenseNet	236 radiograph (cementless implant)
Rouzrokh, Wyles, Kurian, et al. 2022	Acetabular component Inclination	2 Unet + Image processing steps	600 AP pelvic radiograph and 600 Lateral radiograph

Reference	Purpose	Method	Dataset
Kang et al. 2020	implant model (29 types)	YOLOv3 +Simple CNN	170 radiograph (The number of images in each design varied from 2 to 18)
Patel et al. 2021	implant model (12 types)	Unet	922 radiograph (The number of images in each design varied from 42 - 233)
Karnuta et al. 2021	implant model (18 types)	CNN	1972 AP radiograph (The number of images in each design varied from 20 - 345)
Shah et al. 2020	implant loosening	DenseNet	679 Radiograph and demographic data
Loppini et al. 2022	implant failure : loosening, malpositioning,	DenseNet169	1260 Radiographs AP and Lateral THR (840 failed / 420 non-failed)

Table 2.1: Overview of the proposed studies for THR image analysis

2.4 Research Gaps and Limitations

Although the above CNN-based methods provided promising results in the detection of some THR complications, several factors can impact the reliability and generality of these results Firstly, the quality of images used to train CNN-based models can affect the performance and generalization of the model significantly. For instance, Patel et al. 2021 excluded all complex and noisy images from the dataset when generating the ground truth implant masks that were generated using a simple image processing method. In addition, T. Rahman et al. 2022 collected images from different online resources such as medical articles and radiology websites. These images are likely to be the most descriptive and representative of the condition under consideration. Well-represented and clear images may not accurately reflect the actual diversity of the disease presentation and can lead to high-accuracy results as the model learns to recognise particular patterns instead of detecting complications in more challenging and complex images.

In contrast, low-quality and noisy images may minimize the model's ability to learn relevant features leading to lower accuracy and perhaps restricting its usefulness in real-world clinical contexts. Therefore, it is critical to utilize a diverse and representative dataset that reflects the real conditions of the hospital setting.

Secondly, the majority of studies utilized a limited and highly imbalanced dataset, where some classes representing specific types or conditions have significantly fewer samples than others, refer to table 2.1. This might cause a model bias towards the largest classes, resulting in lower performance in the minor classes. Many studies preserved this imbalance in class distributions, whereas Patel et al. 2021 applied class weight and Khosravi et al. 2022 applied oversampling technique, attempting to address this issue. To some extent, these techniques could help reduce the impact of class imbalance but might lead to overfitting. When assigning weight to a small class, the model could focus on it and can not generalize well to new data. This is also applicable to oversampling where generating an exact copy of the small classes leads the model to memorize the training sample and make it less effective in predicting new data.

Another issue of the reviewed work is the limited scope of the task. All the reviewed work on THR complications primarily focused on a detection task rather than identifying the specific type or providing a diagnosis. While detection is a crucial step, providing more detailed information about the specific complication or diagnosis would enhance the clinical utility and practical applicability of the developed models. In addition, no work has been investigated to detect and diagnose PFFs.

In general, the effectiveness of DL-based approaches depends on the availability of large and well-annotated dataset which is considered a main obstacle in several medical image analysis tasks including THR application. It is challenging to generate an adequate large dataset that represents all medical conditions due to the wide range of patients and diseases. In addition, generating annotated medical image dataset requires the involvement of experts which is considered time-consuming and of high cost. Another common challenge is the occurrence of a particular disease which leads to an unequal distribution of classes, resulting in an imbalanced dataset. As illustrated above, many strategies have been applied to overcome the limited amount of the dataset and class imbalance. Many studies considered the transfer learning approach to address the limited dataset size (Rouzrokh, Ramazanian, et al. 2021; Borjali, A. F. Chen, Muratoglu, et al. 2019; Muscato et al. 2023). In addition, augmentation techniques such as rotation, trans-

lation, and flipping the images have been applied to increase the dataset size. Applying these techniques could improve the performance of the DL model in the restricted dataset. Recently, incorporating domain knowledge into DL demonstrated enhancement in model performance and robustness under the dataset limitation issue.

In the next chapter, an annotated THR dataset that consists of images of different implant types with various categories of PFFs and normal THR cases is generated. The performance of different CNN architectures for detecting, localizing and diagnosing the PFF is demonstrated. Chapter 4 constructs an SSM that represents the important landmarks in the implant based on medical protocol. In addition, a hybrid method for the simultaneous detection of the important landmarks and segmentation of the implant components is developed. The defined landmarks can aid in the diagnosis of many THR complications. Chapter 5 demonstrated the impact of incorporating medical knowledge into a DL model for the diagnosis of PFF. The clinical steps for interpreting THR X-ray images are defined and simulated in CNN-based architecture. Chapter 6 focuses on the pre-operative planning for fracture reduction operation. An approach for solving a 3D puzzle problem is developed to reassemble a fractured object into its original form.

2.5 Summary

In this chapter, a medical overview of THR including femur parts, THR components and THR complications is presented. The challenges in THR images such as the quality of the X-ray image and the variations in THR due to many factors are discussed. These factors include surgical approaches, implant design, material selection etc. In addition, an overview of the state-of-the-art approaches for medical image classification and segmentation tasks is provided and the overall challenges in DL models are discussed.

The existing methods for both THR X-ray image analysis and complication detection are reviewed. The main limitation of these methods is defined including the quality of images being utilized, the dataset size, imbalanced class distribution and the limited scope of the undertaken task. Therefore, this thesis develops diagnostic methods for THR-related complications from X-ray images, with a focus on PFF diagnosis. It designs DL-based methods for automated PFF detection and diagnosis, as well as important landmark identification in THR X-ray images which can facilitate the detection of other THR complications. These methods integrate

medical knowledge to enhance diagnostic accuracy and address dataset limitations and image challenges.

Chapter 3

Deep Learning for Implant Joint Fractures

3.1 Introduction

The increasing rate of THR operations is associated with an unavoidable rise in post-operative complications such as PFFs that occur in 4.6% of patients who undergo THR (Abdel et al. 2016). Following a primary THR, PFFs account for 10.5% of revision hip arthroplasties (United Kingdom National Joint Registry 2020). It is usually caused by low-energy falls in elderly patients, but can also be due to implant loosening, osteolysis or stress from an adjacent implant.

The assessment and management of PFFs rely on a clinical assessment of the patient, prior operation notes on the joint implant and surgical approach taken, and it is guided predominately by the associated radiographic appearance to assess the fracture characteristics and the implant status (Ramavath et al. 2020). The management of PFF varies from non-operative treatment to open reduction and internal fixation (ORIF) to revision of the prosthesis or combination of both surgical options based on the fracture type (S. Lee et al. 2019). The Vancouver classification system is a commonly used clinical protocol to characterise these fractures and guide subsequent surgical management (Capone et al. 2017). Figure 3.1 presents the PFFs categories according to the Vancouver classification which categorizes the fracture into three main categories and 6 sub-categories. It considers three main fracture features: fracture location, implant loosening and bone quality.

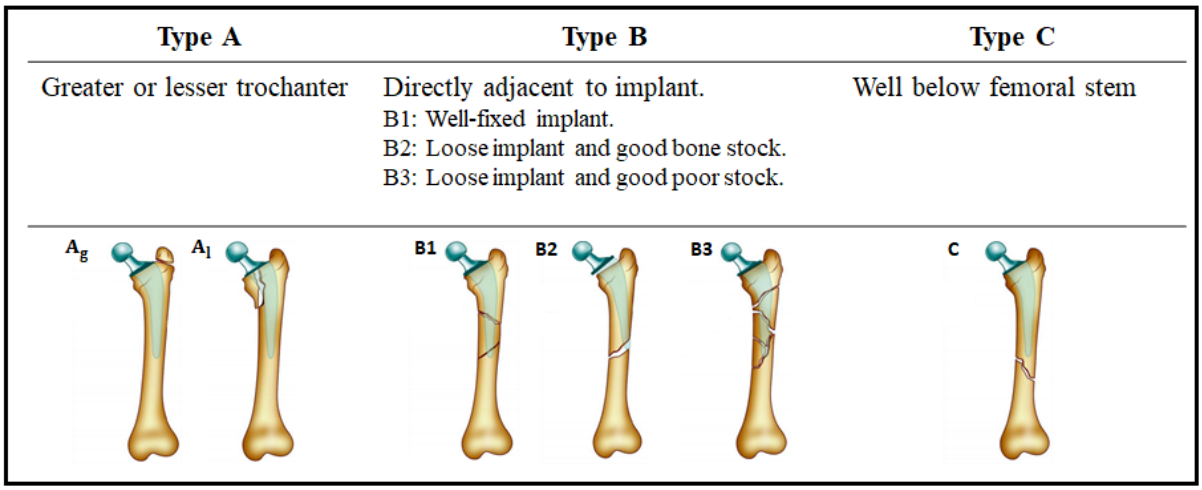


Figure 3.1: The classification of PFFs according to VCS (Schwarzkopf et al. 2013)

The current assessment of PFFs depends on manual examination of the X-ray images. In fact, it has been found that 90% of PFF radiology reports do not include all relevant radiographic features. This may lead to a delay in diagnosis and incorrect treatment strategy and, ultimately delay the surgery (Marshall et al. 2017). Automating the detection and diagnosis of PFF can minimize these problems, lead to better treatment planning and improve patient outcomes.

The huge development of machine learning techniques has had a major impact on improving the detection and diagnosis of different diseases such as Lung nodule detection in the chest (Halder et al. 2020), mass detection and classification into benign or malignant (Houssein et al. 2021). A collection of research and methods on CADx in medical images can be found in (Gao et al. 2019; Debelee et al. 2020). Compared to this development, techniques for automatically diagnosing bone fractures are scarce (Joshi and Singh 2020).

The previous efforts in Computer-aided bone fracture diagnosis tools varied from the detection of the fracture appearance in the X-ray image to the identification of the fracture type. The majority of work focused on the detection of fractures in specific anatomical regions and excluded the fracture cases with a prosthesis (Tanzi et al. 2020). On the other hand, a few existing fracture diagnosis techniques proposed the classification of the fracture type.

The existing methods for automatic image analysis of bone fractures are based either on hand-crafted features or on learning relevant image features. The early work on fracture detection and classification focused on a typical machine learning framework that generally consisted of pre-processing, feature extraction and classification steps. For the pre-processing step, many low-level pixel-processing methods such as noise reduction and segmentation were used to obtain

the region of interest (ROI). Using ROI, various features can be extracted for the classification of bone fractures. The feature types can be texture analysis (J. C. He et al. 2007; Hmeidi et al. 2013; Chai et al. 2011), a combination of texture and shape features (Umadevi and Geethalakshmi 2012) or digital geometry of the extracted fracture points (Bandyopadhyay et al. 2016). For the classification step, the fusion of multiple classifications resulted in improved fracture prediction (Umadevi and Geethalakshmi 2012; Mahendran and Santhosh Baboo 2011) when compared to using a single classification approach (Chai et al. 2011).

The hand-crafted feature-based approaches require prior knowledge of the specific feature to be extracted which affects their generalization ability. In addition, most of these methods relied on a prior segmentation of the bone, a process that typically lacks accuracy in extracting bone contours. Modelling and representing a bone fracture is complex due to a large number of parameters involved but it could be learned from a large set of relevant image data.

The recent developments of DL techniques have overcome some limitations of traditional feature-based approaches. CNNs have demonstrated the ability to detect fractures by performing the binary classification task (fracture or normal) in different anatomical regions, such as hip (Cheng et al. 2019), pelvis (Y. Wang et al. 2019), wrist (Lindsey et al. 2018), spine (H. R. Roth et al. 2016) and ankle (Kitamura et al. 2019). Some approaches leveraged a pre-trained CNN model, specifically trained on ImageNet (Y. Wang et al. 2019; D. H. Kim and MacKinnon 2018; Olczak et al. 2017) or trained on a similar dataset such as bone X-ray images (Cheng et al. 2019; Lindsey et al. 2018) in order to improve the accuracy of the classification and overcome the limitation in dataset size. Moreover, it is illustrated in several studies that cropping the ROI and feeding it to the network increases the classification accuracy (Y. Wang et al. 2019; Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020). Combining hospital process variables such as hospital department, scanner model, and patient demographic information such as age, gender, body mass etc., can further improve fracture prediction outcome when compared to using just X-ray images of the fracture (Badgeley et al. 2019).

All the above studies focused on a specific part of the fractured bone e.g. proximal femur (Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020), and do not consider more diagnostically complex fractures close to joint implants. Figure 3.2 illustrates the challenges of detecting, localizing and classifying the PFF from X-ray images. In addition to the poor quality of X-ray images due to noise and low contrast, there is a wide range of fracture types

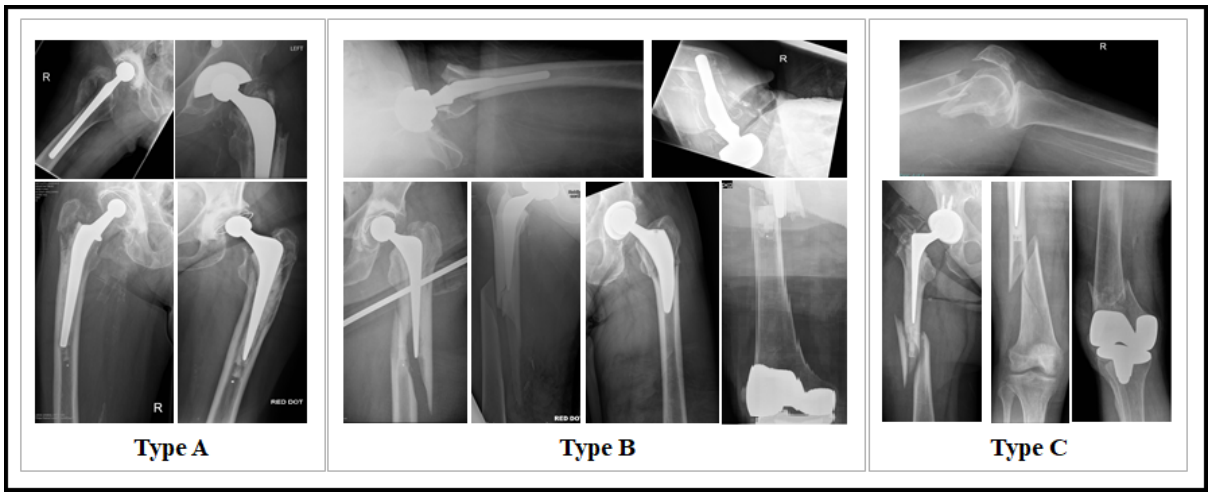


Figure 3.2: Illustration of the quality of X-ray images, fracture line appearance and the high variability in PFFs x-ray images; image view, implant type and captured bone part.

with different visual patterns at different anatomic locations. Furthermore, there is variability in the X-ray images in terms of capturing different parts of the bone for the same fracture type. In contrast to the hip or other aforementioned fractures, which are located at a specific position for example the femoral neck, PFFs can be located anywhere on the femur around or below the implant. This increases the complexity of image pattern analysis and makes the extraction of a ROI based only on the bone anatomy more difficult. Furthermore, the variety of prosthesis types increased the image variations significantly.

This chapter considers in-depth evaluation of different DL approaches and reports results for the detection of the presence of the fracture (binary classification 'fracture, normal'), classification of the fracture according to the Vancouver system and localization of the fracture to tackle the diagnosis of PFFs and assist orthopaedic surgeons in fracture management that can ultimately enhance patient outcomes. For this purpose, a large dataset of PFF images and the annotations of the fracture classes and bounding boxes have been generated for this research.

3.2 Method

In this chapter, a CADx tool based on CNN was developed and different model architectures were systematically explored. In this direction, two approaches were examined: PFF classification (section 3.2.1) and PFF detection, which combined both the classification and localization (section 3.2.2). Figure 3.3 presents a general overview of these approaches.

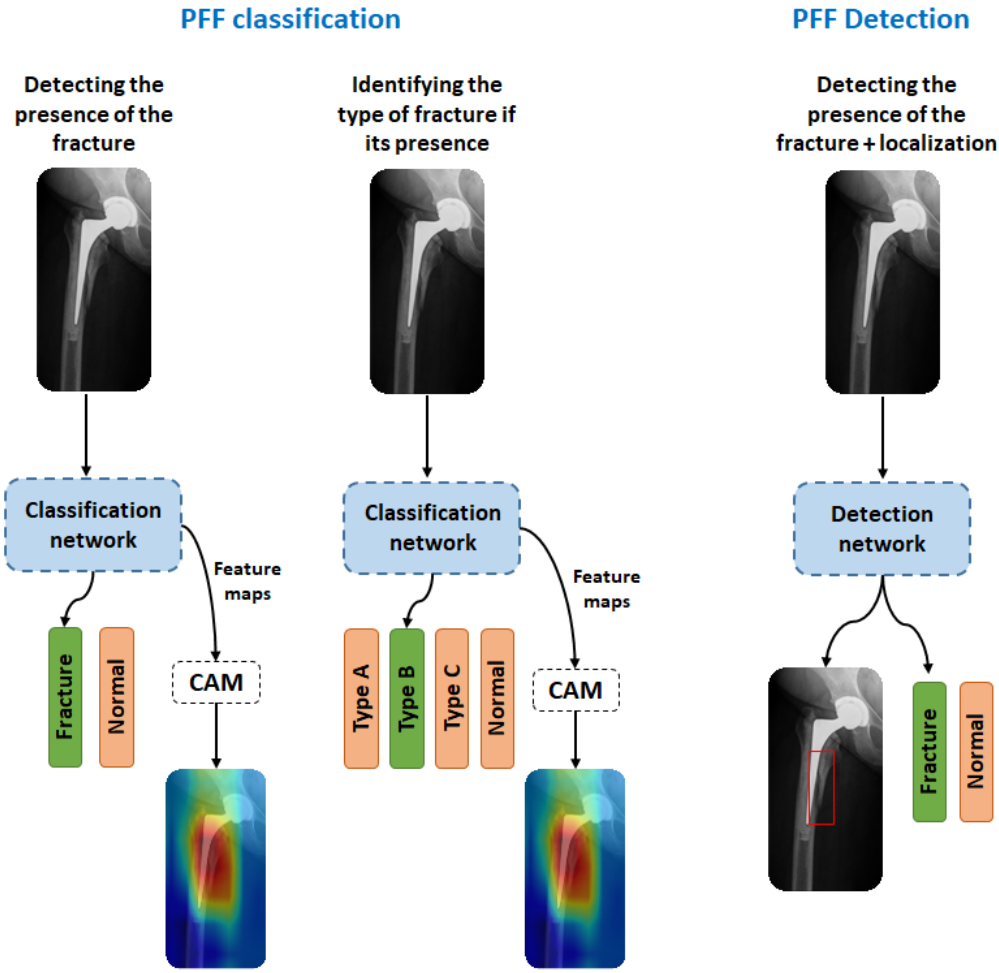


Figure 3.3: PFF classification approaches. The examined classification networks are AlexNet, GoogleNet, ResNet, DenseNet, VGG, ViT and Swin Transformer. The object detection networks are FasterRCNN and RetinaNet.

3.2.1 PFF Classification

Given a set of X-ray images $I \in R^{H \times W}$, the goal was to train a classification model $f(\cdot)$ in order to specify a class label $y \in C$ for each image (I_i). Two sets of class labels were considered - $C \subset$ {*fracture, normal*} for detecting the presence of a fracture and $C \subset$ {*Type A, Type B1, Type B2, Type B3, Type C, normal*} for categorization of the fracture. The classification model can be defined as:

$$y = f(I; w_f) \tag{3.1}$$

Where I is the X-ray image and w_f are the model parameters. The function f is approximated using a CNN optimized to minimize the cross-entropy loss function:

$$\ell_{class} = -\sum_{j \in C} y_{j,c} log(y_{j,c}). \tag{3.2}$$

Visualization of PFFs

The Class Activation Map (CAM) method is used to visualize the fracture region, which generates a weighted activation map for each image (Bolei Zhou et al. 2016). This identified a region that a classification model is focused on. The CAM method depends mainly on the global average pooling layers which are added after the last convolutional layer of the network to create the spatial average of the feature map of each image unit. Given an image, let $f_k(x, y)$ denote the activation of unit k in the last convolutional layer at a spatial location (x, y). Then, the result of average pooling for unit k is $\sum_{x,y} f_k(x, y)$ and the class activation map for class c for a spatial element is defined as:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{3.3}$$

Thus, the class score $S_c = \sum_{x,y} M_c(x, y)$. The $M_c(x, y)$ shows the importance of the activation at (x, y) resulting in the image classification to a class c.

To highlight salient features in the X-ray image that discriminated abnormality, the CAM was up-scaled to the image dimension and overlayed the image.

3.2.2 PFF Detection Approach

The PFF detection approach classified and localized the PFFs using image labels and vertices of a fracture bounding box in a fully supervised fashion. The following sections describe two object detection models: Faster RCNN (Ren et al. 2017) and RetinaNet (Lin et al. 2017).

Faster RCNN

Faster R-CNN is a two-stage object detection model: Region Proposal Network (RPN) and Fast R-CNN. Both stages share the same backbone network, which outputs the feature map of the input X-ray image.

RPN is a Fully Convolutional Network (FCN) responsible for generating region proposals with various scales and aspect ratios which are used by Fast R-CNN for fracture detection. The RPN applies the concept of attention to tell the (Fast R-CNN) where to look. First, a sliding window with a size $n \times n$ is passed through the feature maps to generate K anchors with a different size and aspect ratio for each location. For each pixel, the network checks whether these K anchors contain an object (fracture) or not. Therefore, for each anchor, a feature vector is extracted and fed to two fully connected layers. The first one is a binary classifier that computes the objective score i.e if the area includes an object (fracture) or not. The second one returns the bounding box as region proposals.

Fast R-CNN: The feature maps from the backbone network and the resulted region proposals are fed to the ROI pooling layer. The ROI pooling layer splits each region proposal into grid cells and applies a max pooling operator to each cell to return a single value. The output feature vector is defined by all values from all these cells. The feature vector is then passed to the fully connected layer which is divided into two sub-networks: the softmax layer that predicts class scores and the regression layer that predicts the bounding box coordinates.

RetinaNet

RetinaNet is a one-stage object detection model, which consists of three sub-networks: a backbone network, a Feature Pyramid Network (FPN), and FCNs.

Backbone network: computes a feature map of the input X-ray image.

FPN: is used to construct a rich multi-scale feature pyramid from a single-scale input image.

The structure of the pyramid consisted of two pathways: bottom-up and top-down. The first pathway computes a feature hierarchy by using the feature activation output of each residual block. The high-level feature maps are considered in the top-down pathway by up-sampling spatially coarser feature maps from the higher pyramid levels.

FCNs: This sub-network includes two FCNs. The first FCN performs the classification task (fracture/ no fracture), while the second one performs the bounding box regression (localization of the fracture).

RetinaNet uses a focal loss function to resolve the class imbalance problem between the background and foreground in the detection scenario. Thus, the standard cross entropy loss has been modified to the following:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t). \tag{3.4}$$

where $p_t = \begin{cases} p & \text{if y=1} \\ 1-p & \text{if otherwise} \end{cases}$

γ is a tuning focusing parameter ($\gamma \geq 0$)

3.2.3 Dataset Collection and Preparation

The dataset of PFF images was collected at multiple trauma centres in the United Kingdom. Overall, 607 anonymised patient data were collected with a total of 2544 X-ray images. To establish a ground truth classification and detection for the images, two clinical experts participated in image annotations and provided class labels and fracture bounding boxes.

For each patient, either a lateral or an anterior-posterior (AP) image or both was collected. The images included either a partial region of the femur, the full femur or the pelvis with both femurs. The last type of image was split into two, containing one femur each. The images were of various scales, orientations and implant types. Images for each patient included an X-ray after THR surgery (representing the normal cases) and an X-ray containing the fracture. The images were annotated to Fracture and Normal. Part of the fractured image samples were annotated by fracture type (Type A, Type B1, Type B2, Type B3, Type C). The fracture images were further annotated by a bounding box around the fracture i.e. the coordinates of

the minimum and maximum corners of the rectangle. Microsoft Visual Object Tagging Tool (VOTT) was utilized for image annotations.

PFF classification: For the classification task, both binary classification (fracture vs normal) and multi-classification (Type A, Type B1, Type B2, Type B3, Type C and normal) were considered. For binary classification, 1272 images with a fracture and 1272 images without a fracture (normal) were used.

In the multi-classification task, the number of normal images was very high when compared to the other types, therefore, random images were selected from normal image samples. The dataset consisted of 70 normal, 48 Type A, 80 Type B1, 87 Type B2, 37 Type B3 and 67 Type C images. For both tasks, the dataset was divided into two parts: training and validation, with the ratio 75% : 25%, respectively.

PFF detection: The same dataset of fracture images in the binary classification experiment was used in PFF detection and split into the training and validation sets. The boundary box of the fracture region was represented as the coordinates of the upper left and lower right corners of the rectangle. Two classes were considered i.e. fracture and background (normal).

3.3 Experiments

3.3.1 Model Architectures and Implementation Details

All the models were trained on a Windows machine equipped with 8 GB RAM, Intel(R) Core(TM) CPU @ 3.00 GHz and GeForce RTX 2080 graphics card.

PFF classification: For classification tasks, seven CNN architectures were compared: AlexNet (Krizhevsky et al. 2012), GoogleNet (Szegedy et al. 2015), ResNet50 (J. C. He et al. 2007), VGG (Simonyan and Zisserman 2015), DenseNet121 (G. Huang et al. 2017), Swin Transformer (Z. Liu et al. 2021) and ViT (Dosovitskiy et al. 2020). All of these models were pre-trained on ImageNet. Each network was trained on X-ray images down-sampled from the original size to 224 × 224 px, except AlexNet model which was trained using 256 × 256 px image size. The classes included 'normal' and 'fracture' for the binary classification and 'normal' and the categories of the Vancouver system for multi-class classification. Data augmentation techniques such as flipping, rotation and scaling were used. An experiment utilizing copy-paste augmentation was conducted to augment the dataset; however, it did not lead to an improvement in the results.

The CAM was used on top of each model to visualize the fracture region.

Stochastic Gradient Descent (SGD) was used for optimization. All the models were trained until convergence (100 epoch). The batch size was 8, momentum 0.9 and the learning rate was set to 1 $\times 10^{-2}$.

PFF detection: Both models were trained and validated using different image resolutions. For the backbone network, ResNet50 was used in both object detection models and the optimization was performed using SGD. All the models were trained until convergence (100 epoch). The batch size was 2, Momentum was 0.9. The default anchor configuration and non-maximum suppression with Intersection Over Union (IOU) 0.7 was used. The learning rate was set to 1 $\times 10^{-2}$ on Faster R-CNN and 5 $\times 10^{-2}$ on RetinaNet. For the focal loss function, the default value for γ as provided in PyTorch was used.

3.3.2 Evaluation Settings

To evaluate the classification results, standard metrics derived from Confusion Metrics were applied including accuracy, precision, recall (sensitivity), specificity and F1 score. The balanced accuracy is used for evaluating the multiclass classification due to the imbalance distribution of the classes. The binary classification accuracy determines how many observations, both positive and negative, were correctly classified, it is defined as:

$$Accuracy = \frac{TP + TN}{N} \tag{3.5}$$

where TP is the number of true positives, TN is the number of true negatives and N is the total number of samples. The precision measures the proportion of predicted positives that were actually correct, it is defined as:

$$Precision = \frac{TP_k}{TP_k + FP_k} \tag{3.6}$$

where TP_k is the number of true positives in class k and FP_k is the number of false positives in class k. The recall measures the proportion of actual positive samples that were identified correctly, it is defined as:

$$Recall = \frac{TP_k}{TP_k + FN_k} \tag{3.7}$$

where FN_k is the number of false negatives in class k. F1-score computes the harmonic mean of precision and recall to summarise the overall performance of the classifier. It is defined as:

$$F1score = \frac{2 \times Precision \times Recall}{Precisison + Recall} \tag{3.8}$$

Specificity measures the proportion of predicted negative samples that were actually correct, it is defined as:

$$Specificity = \frac{TN_k}{TN_k + FP_k} \tag{3.9}$$

Where TN_k is the number of true negatives in class k.

The balanced accuracy is the mean of sensitivity (recall) and specificity which is defined as:

$$Balanced_accuracy = \frac{Sensitivity + Specificity}{2} \tag{3.10}$$

In addition, the part of the X-ray image that contributes more to the prediction was visualized as explained in section 3.2.1 for a qualitative analysis of the clinical applicability of the classification model.

For the object detection task, the localization accuracy is measured which considered the tested image as correct if both predicted classes and the bounding box were correct. The correct bounding box was defined using the IOU measure which computes the overlap area between the ground truth box and the predicted box over the area of the union of them. The predicted bounding box was considered correct when IOU ≥ 0.5. In addition, the precision, recall and Average Precision (AP) are reported. The precision is defined as in Equation 3.6 where TP is the number of correct detection made by the model i.e. IOU ≥ 0.5. FP is the number of incorrect detection made by the model i.e. IOU ≤ 0.5. The recall is defined as in Equation 3.7 where TP is the number of correct detection made by the model i.e. IOU ≥ 0.5 and FN is the number of missing detection of objects. Average Precision is defined as the area under the precision-recall curve.

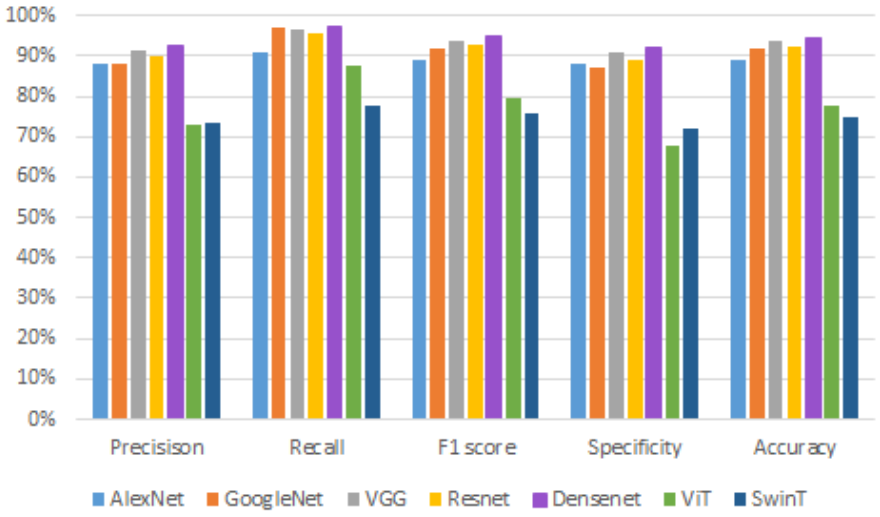


Figure 3.4: Comparison of the performance of Fracture/ no fracture classification. Swin Transformer and ViT models exhibited lower metric values compared to others. Minor performance variations were observed among the remaining models, with consistent results across metrics. DenseNet outperformed the others.

3.4 Results

3.4.1 PFF classification

Two main classification experiments of PFFs were evaluated - binary classification to distinguish between fracture and normal X-ray images and classification according to Vancouver classification system.

Figure 3.4 presented the accuracy, precision, sensitivity (recall), specificity and F1 score for each evaluated model of the binary classification task. Notably, the Swin Transformer and ViT models showed lower values across these metrics compared to the other models. Comparing Swin Transformer and ViT, ViT achieved better performance in detecting fracture samples with a Recall of 88%, F1 score of 80% and accuracy of 78%. On the other hand, when comparing the remaining models, minor performance variations were observed, with consistent results across the evaluated metrics. In terms of Precision, Densenet121 achieved 92%, while VGG and Resnet showed slightly lower precision with 1 % and 2% differences, respectively. On the other hand, AlexNet and GoogleNet were 4% lower compared to Densenet121. In regards to correctly identifying fracture samples, all the remaining models provided outstanding outcomes, with Densenet, VGG and GoogleNet achieved a recall of 97% and Resnet achieved 96%, whereas AlexNet achieved 91%. Densnet showed the highest F1 score of 95%, closely followed by VGG,

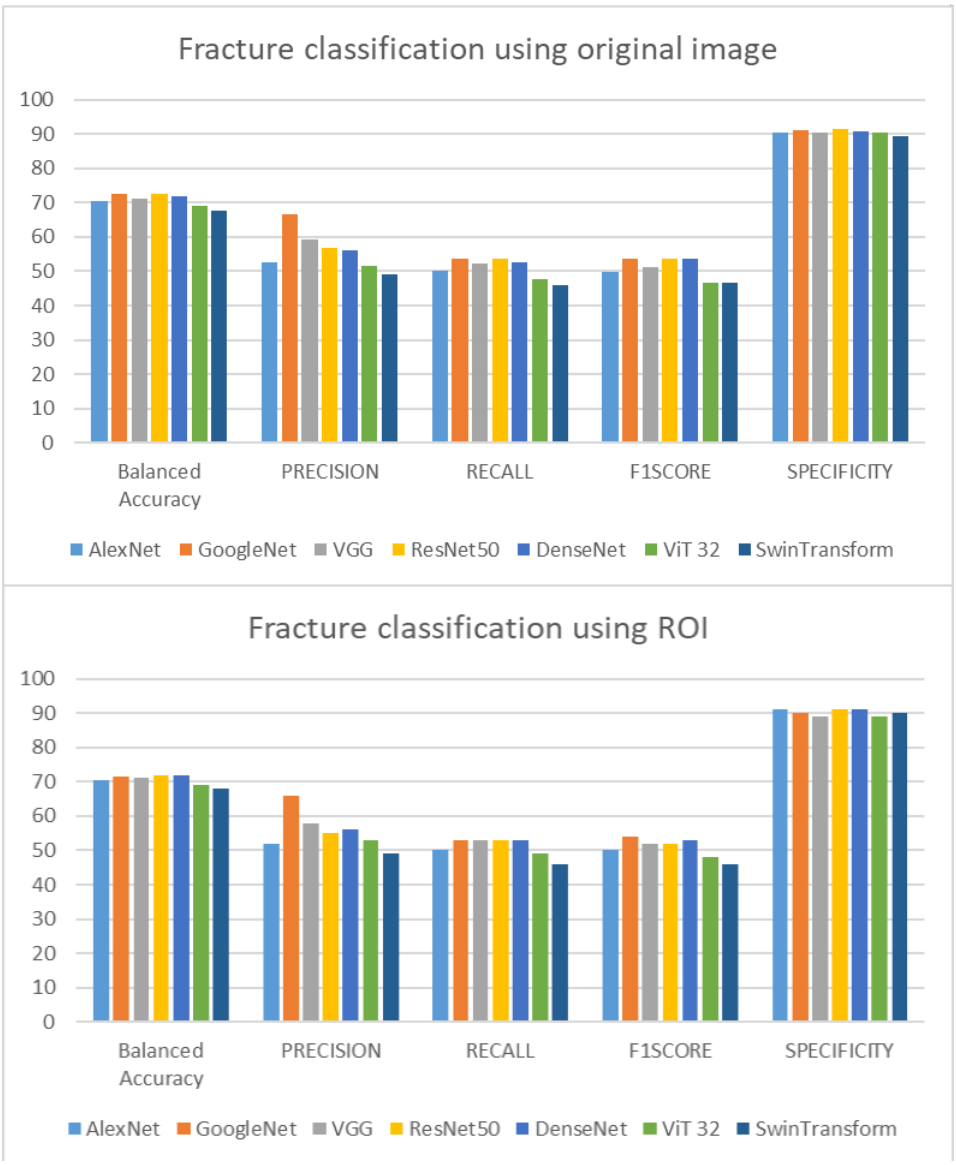


Figure 3.5: Balanced Accuracy, precision, recall, F1-score and specificity of PFFs classification using original image and using ROI.

Resnet and GoogleNet with a 1%, 2% and 3% difference, respectively, while AlexNet provided the lower F1 score of 89%. Similarly, for accuracy, Densenet achieved the highest accuracy result of 94% indicating its outstanding performance among all the models in detecting PFF.

For classification based on Vancouver types (multi-class classification task), Figure 3.5 demonstrated the balanced accuracy, precision, sensitivity (recall), specificity and F1 score for each model. Additionally, the figure showed the results of using the full image as input and using the cropped ROI i.e. femur region, to assess the impact of using ROI on the classification performance. The figure clearly demonstrated a decline in performance compared to binary classification as the task complexity increased, particularly observed in the F1-score and Ac-

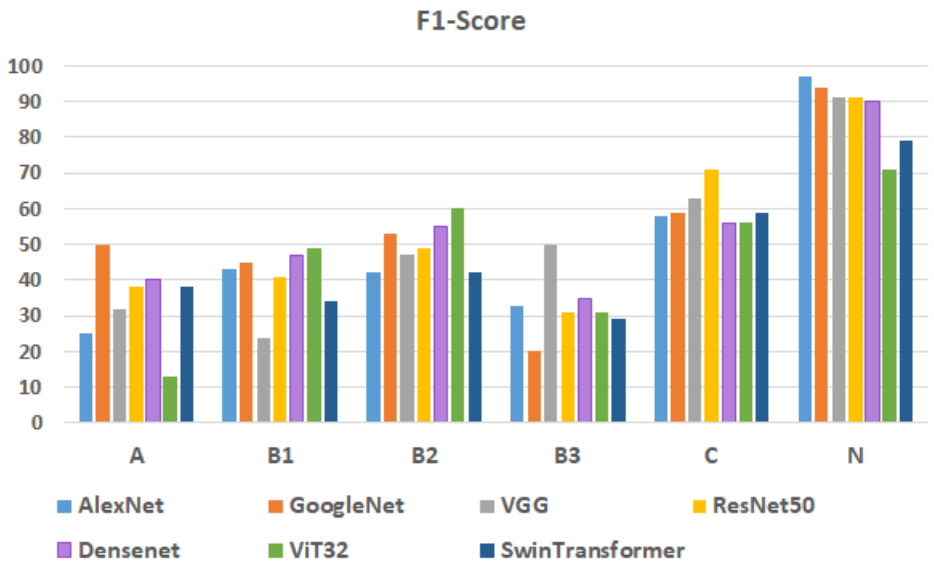


Figure 3.6: Comparison of classification models for normal and each fracture types. The performance of the models varied across different classes, showing the best in classifying Normal images while demonstrating low performance in identifying B3 fractures.

curacy, which decreased by 40% and 10% respectively. It is also observed that incorporating ROI images resulted in similar performance as that of the full image, which contrasted to the findings of other fracture classification methods such as in Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020 method, where ROI significantly improved the classification results. Among the evaluated models using the full X-ray image, both GoogleNet and Resnet models achieved the highest scores across multiple metrics with a balanced accuracy of 73%, recall and F1 score of 54% and specificity of 91%. However, DenseNet, AlexNet and VGG demonstrated slight variations in performance. DenseNet resulted in a similar F1-score and 1% lower in the balanced accuracy and recall metrics. VGG achieved 71%, 52% and 51% while Alexnet achieved 70%, 50% and 55% in balanced accuracy, recall and F1 score, respectively. On the other hand, the Swin Transformer and ViT models illustrated the lowest values across all metrics compared to the other models with F1 score of 47% for both models and balanced accuracy of 69% for ViT-32 and 68% for Swin transformer.

Figure 3.6 displayed the comparison of F1 score for each class using the evaluated models. The bar chart showed variations in F1 scores among different classes, indicating the models' performance was not consistent for all classes. The normal class achieved the highest values, ranging from 71% to 97%, highlighting the ability of models to correctly identify normal cases. Following, Class C provided the second-highest F1 score within the range of 50% to 71%,

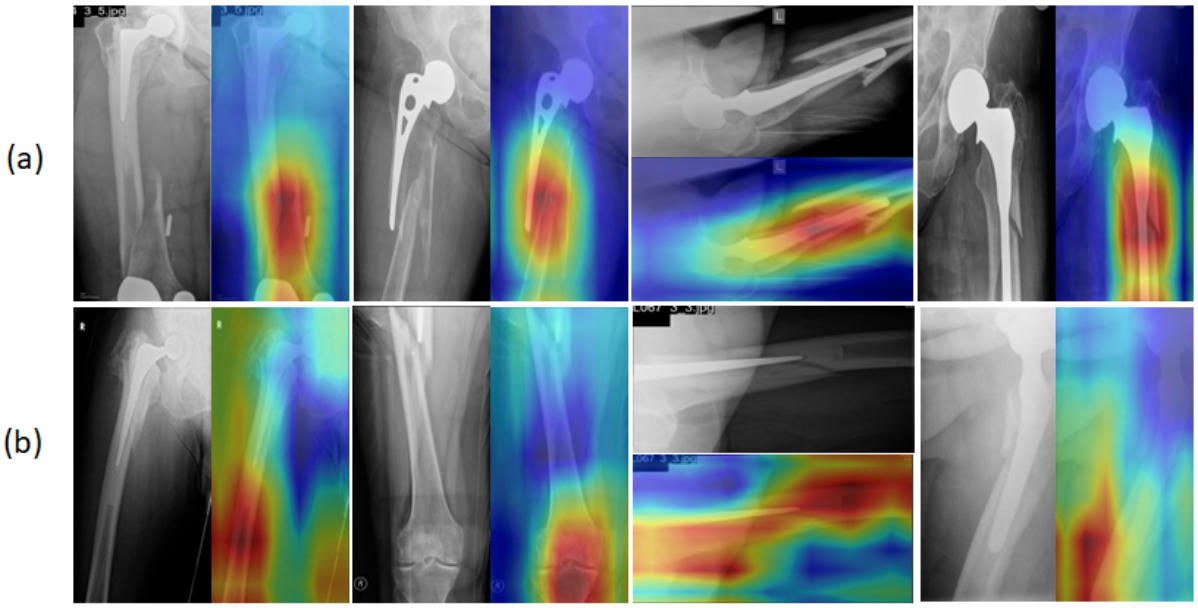


Figure 3.7: Examples of using CAM method for ResNet50. (a) visualization of correct classification. (b) visualization of incorrect classification. The heat map colour ranges from blue (minimum) to red (maximum).

illustrating its relatively good performance in classification compared to other classes. Class A and B3 provided the lower F1 score among the majority of the assessed models, with the highest score reaching only 50%. This highlights the difficulty of distinguishing these types of fractures. On the other hand, classes B1 and B2 showed slightly better performance in type identification.

Among the assessed models, ResNet demonstrated one of the best classification results. Therefore, the region that the Resnet model focused on to predict a class type is highlighted in Figure 3.7. Part (a) of the figure showed the correct class visualization, it is observed that the fracture region of the image contributed most to the ResNet decision for prediction. On the other hand, part (b) showed incorrect predictions and the model was unable to focus on the fracture regions in these images.

3.4.2 PFF Detection

This section demonstrated the results of two state-of-the-art object detection models Faster R-CNN and RetinaNet for localizing fractures in X-ray images. Table 3.8 presented the precision, Recall and accuracy obtained by the two detection models. As illustrated in the table, Faster R-CNN outperformed the other model, indicating its outstanding performance. The recall results highlighted that both models were able to detect the majority of fractures in images. However,

Method	Precision (%)	Recall (%)	Accuracy (%)
Faster-RCNN	80	98	78
RetinaNet	31	97	31

Figure 3.8: Precision, recall, and accuracy of PFFs detection (classification and localization)

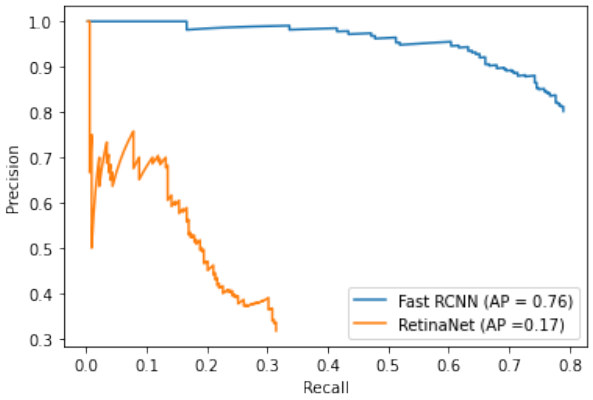


Figure 3.9: Precision-Recall curve for Faster RCNN and RetinaNet

the precision results demonstrated a significant difference between the two models, with Faster R-CNN achieving 80% correct detections compared to only 31% for RetinaNet. Additionally, the localization accuracy of Faster R-CNN, which reached 78%, outperformed the RetinaNet model.

Furthermore, Figure 3.9 analysed the performance of the two models by considering the precision-recall curve. It is evident that Fast R-CNN provided the best performance with an Average Precision value of 76 compared to RetinaNet which achieved a very low Average Precision of 17.

Figure 3.10 presented examples of the predicted fracture locations using Faster R-CNN, RetinaNet and boundary boxes inferred from CAM, alongside the ground truth boundary boxes for comparison. The figure illustrated that RetinaNet provided incorrect boxes in normal images. On the other hand, Faster RCNN demonstrated better localization accuracy, with its detected fracture boundaries being closer to the ground truth, indicating its ability in fracture detection.

3.5 Discussions

PFF is becoming increasingly common as a complication of THR, causing the need for revision hip arthroplasty. The diagnosis and treatment planning for PFF are complicated and depend mainly on the clinical interpretations of femur X-ray images. This requires experienced radiologists who are familiar with the clinical protocol used to evaluate such fractures, however, the majority of PFF radiology reports lack important radiographic details, leading to delayed diagnosis and ultimately impacts on the overall treatment(Marshall et al. 2017).

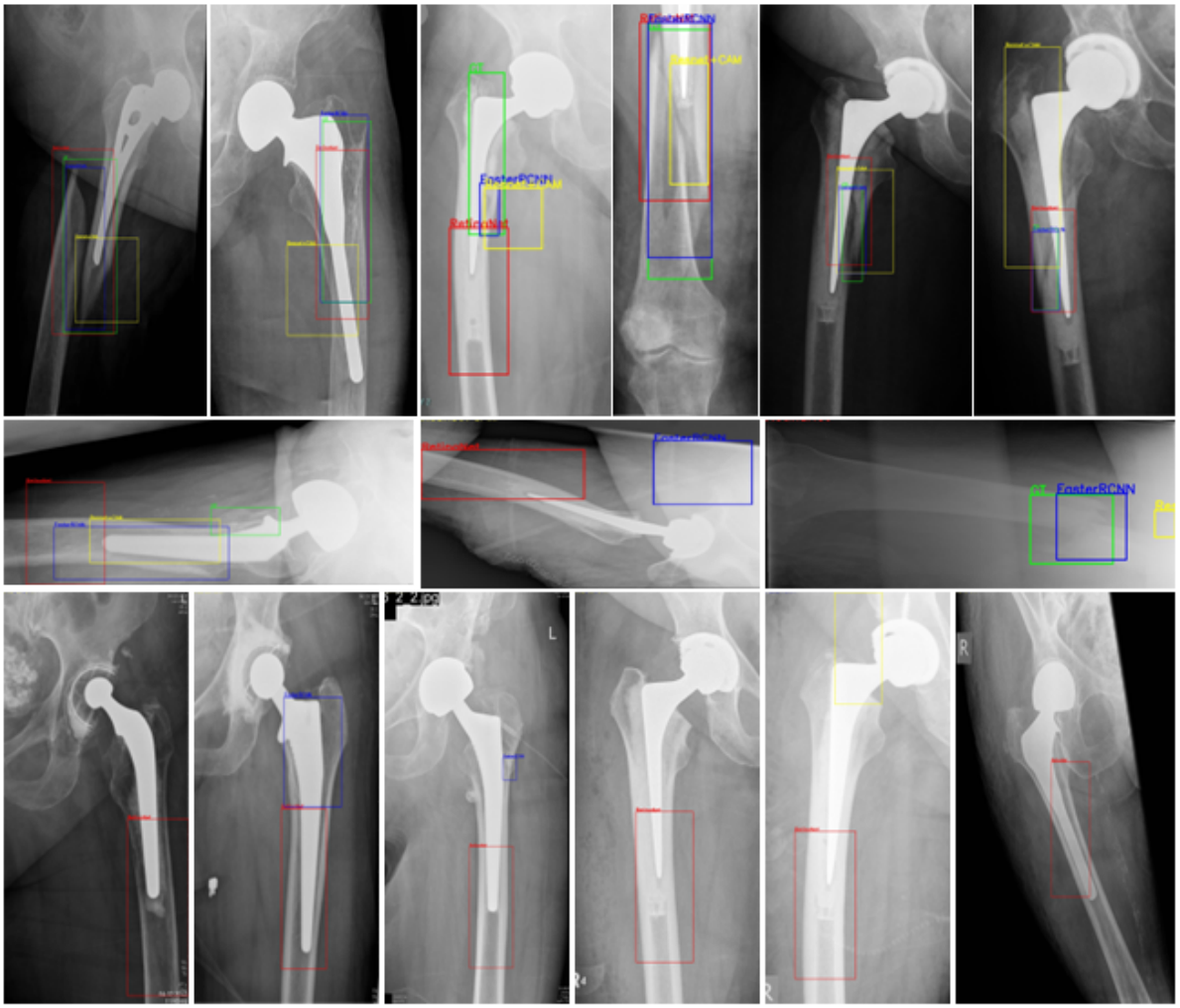


Figure 3.10: The detection of fracture results using Faster RCNN (blue), RetinaNet (red) and boundary box computed from CAM (yellow). The ground truth boundary box (green), when there is no fracture (normal cases) no green box.

In the medical domain, DL models have demonstrated successful outcomes with the aid of different disease diagnoses such as detecting Pneumonia and lung cancer in chest X-ray images. DL models have also shown promising results in detecting bone fractures, such as wrist, hip, and ankle fractures, from X-ray images, whereas the research on fracture diagnosis remains relatively limited compared to other diagnostic applications. Additionally, many fracture detection research excluded fracture cases with a prosthesis. To tackle this limitation, the presented work aims to assess the effectiveness of DL models in the detection and diagnosis of PFF and generates a dataset to support the research in this field. Different DL models were examined for multiple tasks including fracture detection, fracture diagnosis and fracture localization. The typical DL models provided outstanding performance in the detection of fracture (the binary classification task), in contrast to the diagnosis of the fracture type (the multi-classification task), which showed a significant decrease in the overall performance.

In the binary classification task, DenseNet121 achieved the best performance among the assessed models with an accuracy of 95%, demonstrating its effectiveness in detecting PFF. GoogleNet, VGG and Resnet also showed competitive results, with accuracies ranging between 92% to 94%. On the other hand, Swin Transformer and ViT provided the lowest detection results. The comparatively lower performance of Swin Transformer and ViT can be possibly caused by the challenge of working with a small dataset. Large datasets are generally necessary for transformer-based models, such as Swin and ViT, to efficiently learn complex patterns and representations. The small amount of the dataset may have restricted these models' ability to generalise well in this situation.

As no existing automated methods for PFF detection were found in the literature, the achieved result can be considered the state of the art. In comparison to Miao et al. 2019 method for detecting fractures in the femur without the presence of implants, the achieved results outperformed their stated accuracy of fracture detection which was 91%.

In the multi-classification task, Googlenet and Resnet provided the highest results among the assessed models with an accuracy of 86%. An important aspect to consider when evaluating the diagnosis method is how well the model correctly identifies the fracture type. This can be effectively demonstrated by the F1 score, especially in an imbalanced dataset. Notably, the performance of all models showed a significant decline in the F1 score compared to the binary classification task, with a decrease ranging from 29% to 43%. This could be related to the

increased level of complexity, where patterns in binary classification are more distinguishable compared to patterns required to discriminate the Vancouver system classes. Such complexity is reflected in the performance analysis of the models among the defined classes, refer to Figure 3.6. Notably, normal images can be classified more accurately than other fracture types. Also, fracture type C has better distinguishability compared to fractures types A, B1, B2, and B3. This difference can be due to the location of the fracture. While fracture C is located well below the implant, fractures B1, B2 and B3 are located in the same regions, leading to increased complexity in distinguishing these types. Furthermore, the dataset used for multiclass classification was relatively small and imbalanced in comparison to binary classification.

The existing fracture diagnosis approaches suggested that using the ROI resulted in better classification performance. For instance, Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020 applied an ROI cropping method to localize the proximal femur region in a pre-processing step of fracture classification. This allows the variety between the images to be reduced and the model to learn the shape of the proximal femur. However, in PFFs the fracture can be located at different regions of the femur. In addition, the analysed X-ray images contained different regions of the femur which further increased the image variation. Therefore, the classification of PFFs using a femur region as an ROI had a similar accuracy as when the full image was used as shown in Figure 3.5.

The presented work also demonstrated the result of the classification model by visualizing the region that the model paid more attention to in the classification of the fracture using CAM method. The CAM method provides only an approximate localization of the fracture because it tends to concentrate on the most discriminated region of the fracture. A bounding box of the fracture is extracted using the heatmaps computed from the CAM method by firstly thresholding the heatmap image. Then, a list of contours is defined using Suzuki et al. 1985 method. The largest rectangle is selected by defining all the up-right bounding rectangles from the contours list.

A weakly supervised object detection approach, such as the CAM-based method, utilized image-level labels only to classify and localize fractures in the images. On the other hand, the fully supervised object detection approaches used both image labels and fracture region annotations in the training phase. Therefore, the performance gap between the two approaches is still large (Shao et al. 2021). The localization of PFF fractures in X-ray images can be difficult to narrow

to the boundary box so the box may include multiple anatomical regions. This increases the ambiguity of the bounding box. However, the Faster R-CNN provided promising results for PFF localization (Figure 3.10).

3.6 Summary

There are increasing cases of PFFs in the elderly population, associated with the increase in rates of THR. An accurate clinical diagnosis for this type of fracture is essential for taking a correct treatment approach and, subsequently, for the overall clinical patient outcome. Unlike existing techniques developed for fracture detection, this work concentrates on a framework for automated diagnostics of fractures in the proximity of joint implants (hip). An in-depth evaluation of different CNN models for fracture detection, diagnosis based on medical standards and fracture localization. To this end, a large dataset has been collected and annotated with fracture labels and a bounding box.

The results demonstrated that DenseNet is able to detect PFFs with an F1 score of 95%. On the other hand, the classification of the fracture type showed lower performance with an F1 score of 54% achieved by GoogleNet, Resnet and DenseNet. The results also illustrated that using the ROI region achieved similar classification performance compared to using the full image. This is contrasted with the existing approaches for fracture diagnosis such as proximal femur fracture in (Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020) where using ROI reduced the variety between the images and improved the performance of the model. The CAM method provided an approximate visualization of the fracture region. However, Faster RCNN predicted a narrower bounding box of the fracture region with a localization accuracy of 74.5%.

In conclusion, the development of a CADe tool for PFF using CNN-based models like DenseNet or ResNet provided effective and outstanding performance. However, as the task complexity increased, particularly in diagnosing different fracture types, these typical CNN models faced challenges, especially with dataset size limitations and class imbalance challenges. Nevertheless, these models did show promising results for the development of such a tool.

Chapter 4

Simultaneous Detection of Gruen Landmarks and Segmentation of Implant

4.1 Introduction

THR follow-up radiographs are used in routine prosthetic joint evaluation and monitoring to identify any potential complications. These include loosening, infection, and other short and long-term problems related to the region surrounding the implant. For instance, Aseptic loosening, which is the most common cause for THR revision (United Kingdom National Joint Registry 2020) is detected by visually assessing the radio-lucencies 'gaps' around the implant and determining the implant's positional variations in relation to the bone. The widely used clinical protocol for assessing the implant status is the Gruen zone system, which divides the interface between the bone and implant into seven zones (see Figure 4.1 (a)). In clinical practice, these landmarks and the surrounding boundary of the implant are defined by clinicians, often time-consuming and prone to human error processes that could lead to inconsistencies in outcomes between various clinical specialists. Automating the identification of these landmarks and segmenting the implant can minimize

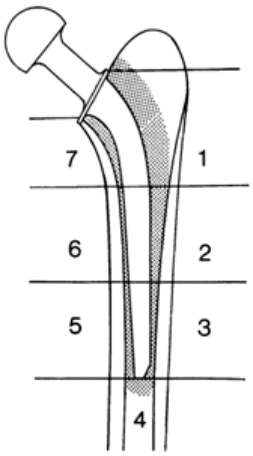


Figure 4.1: Femoral component zones (GRUEN et al. 1979).

these problems and ultimately lead to more efficient and reliable diagnoses, better treatment planning, and, ultimately improved patient outcomes.

In several research studies (Barker and Donnelly 2003, Al-Zadjali 2017) and medical imaging and analysis-assisted tools in orthopaedics such as Ortho View and ELBRA, manual selection of anatomical landmarks or implant boundaries are used for subsequent analysis. There is currently no existing work on automated identification of the Gruen landmarks. On the other hand, several studies attempted to automate the segmentation of hip implant. The early work on implant segmentation considered the analysis of images based on hand-crafted features such as histogram thresholding (Oprea and Vertan 2007, L. Florea, C. Florea, et al. 2011), Active Contour method initialized by using the Fast randomized circle detection method (Al-Zadjali 2017), and the region growing method initialized by applying the Hough transforms (Stark 2018). These methods are not generalized well towards THR radiographs and could provide good results only when the implant components are clearly presented in the X-ray images. Moving from traditional-based methods to DL-based methods, Patel et al. 2021 applied U-Net to segment hip implants as an initial step for classification of the type of implant. Even though CNN showed state-of-the-art results in many medical segmentation tasks, these networks map the global shape structure and can not define the local regional properties. In addition, these networks could produce unrealistic segmentations i.e. gaps or missing parts in the segmented implant, especially when the training dataset is limited, which is considered a major challenge in many medical imaging research. Similarly, the Gruen landmarks might not have simple distinguishable features to be learned by a CNN. It is defined based on the shape and geometry of the implant and its surrounding bone. CNN exceeds at learning hierarchical representations of the visual features but may have difficulty capturing precise geometric features and shapes, particularly if the training dataset is small.

Increasing the dataset size would improve the performance of CNN-based methods, however, it is difficult and time-consuming to annotate a large number of THR X-ray images. In addition, the quality, complexity and variety of THR images may limit the effectiveness of synthesizing new data (Skandarani et al. 2023). Therefore, this chapter introduces a hybrid approach that leverages the shape knowledge of hip implants for simultaneous segmentation and detection of Gruen landmarks in the implant. Although several studies in the medical image analysis domain incorporate shape knowledge into DL such as segmentation of left ventricle (Medley

et al. 2019), brain boundary (Nguyen et al. 2022) and skin lesions (Mirikharaji and Hamarneh 2018), this is the first work that uses such an approach for implant shape segmentation and landmark localization.

There are many currently adopted approaches in the medical image domain that introduced the integration of shape knowledge with CNN. These approaches can be divided into five main categories: (1) post-processing by shape model, (2) prior knowledge, (3) multiple CNNs and shape models, (4) learning hidden representations of shape and (5) shape prior as regularization in the objective function.

In the first category, the shape model was used as a post-processing step to refine the CNN segmentation. Xing et al. 2015 introduced a method for nuclei segmentation that initialized the segmentation with a CNN-based model which generated probability maps of the image. Followed by a selection-based sparse shape model and a local deformable model to perform the final segmentation. Medley et al. 2019 proposed a modified Active Shape Model (ASM) to refine the segmentation of the left ventricle. Since the main limitation of ASM is the searching for landmarks which potentially results in high outliers, the authors took advantage of CNN which maximized the quality of features extraction from images. Then, the Expectation-Maximization was used to minimize the effect of outliers during the ASM optimization. Rather than using segmentation maps for initializing the shape model, Tabrizi et al. 2018 predicted the bounding boxes as initializations, and then produced the final segmentation using the weighted fuzzy ASM. A similar approach was introduced in Y. Li et al. 2017 for myocardial segmentation, however, they applied random forest to build probability maps from the detected bounding box. Then, SSM was utilized for the final segmentation.

In the second category, shape knowledge was applied to generate the initial segmentation. Nguyen et al. 2022 divided images into different groups that have similar shapes and structures of brain boundaries. Then, prior ASM for each group was used to generate coarse segmentation that was followed by a CNN and post-processing methods such as Conditional Random Field and Gaussian processes to refine the segmented contours. Zotti et al. 2018 extended the U-net architecture by incorporating multi-resolution input and integrating a shape prior as a template for cardiac MRI segmentation. A shape priors encoded the probability of a voxel being part of a specific class. The shape prior was used in segmentation and in predicting the centre location of the object.

The approaches in the third category attempted to extract more information from data and provided more accurate outcomes by involving multiple CNNs and shape models. Ambellan et al. 2019 proposed a pipeline that consisted of multiple CNNs and SSMs to segment knee bone and cartilage from MRI images. The pipeline started with 2D U-Net to generate initial segmentation masks which were then regularized by SSMs. Then, 3D U-Net was employed to extract smaller MRI subvolumes. To enhance the results, another SSM was used as a post-processing step. In the end, a third U-Net was used to segment the cartilage. Similarly, Brusini et al. 2020 proposed three steps pipeline for segmentation of the hippocampus from brain MRI. It consisted of U-Net, SSMs and a second U-Net. However, they utilized three orthogonal U-Nets and averaged their prediction to extract the final segmentation. J. Duan et al. 2019 developed a segmentation method for cardiac images that combined a multi-task DL approach with an atlas as a prior shape. Their method trained FCN for both segmentation and landmarks detection. The landmarks were used to initialize the atlas by selecting the most corresponding one which was used to refine the segmentation.

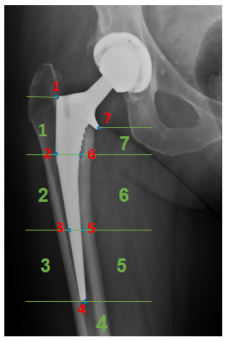
The fourth approach is based on the hidden representation of the shape. Larrazabal et al. 2019 proposed denoising autoencoders (DAE) as a post-processing step for lung segmentation. The DAE was trained using the segmentation masks and utilized the learned representations of the anatomical shape and topological structures to impose the shape constraints on the initial segmentation. Another approach (Painchaud et al. 2020) developed a constraint variational autoencoder (cVAE) for learning the latent representation of cardiac shapes. Then, the post-processing VAE encoder was used to generate a shape vector from the initial segmentation which was constrained by the latent shape space to correct the shape and passed to the VAE decoder to generate the final segmentation. C. Chen et al. 2019 proposed a Shape-aware Multi-view AutoEncoder (Shape MAE) for learning the anatomical shape priors of cardiac. Then, a multi-view U-Net was used to perform the segmentation by integrating the learned anatomical priors into the feature maps of the segmentation.

The last category includes the approaches that combined the shape priors as regularization terms in the loss function of the segmentation network. There are two ways for defining this loss; either by using the landmarks distance (C. Qin et al. 2022, Schock et al. 2020) or based on the shape parameters (Karimi et al. 2018, Tilborghs et al. 2020). For instance, C. Qin et al. 2022 developed two parallel networks for prostate segmentation. The first one was an inception-

based network to predict the SSM parameters that can be converted to prostate contour. A normalized distant map was then computed as an output of this branch. The second branch was a Residual U-Net that generated the probability maps from the input image. The two outputs were combined to generate the final segmentation. Karimi et al. 2018 proposed a CNN-based regression model for predicting the centre of the prostate and the shape parameters of the shape model. A stage-wise training process was adopted which trained the network for predicting the centre location first, then, added the shape model parameters and rotation vector to the network. A similar approach was introduced for left ventricle segmentation which performed a regression of shape and pose parameters (Tilborghs et al. 2020), however, they integrated the regression of distance maps as a third output to perform the semantic segmentation. (Mirikharaji and Hamarneh 2018) encoded a star shape prior as a new loss term in an FCN to enhance the segmentation of skin lesions. To guarantee a global structure in segmentation outcomes, the non-star shape segments were penalized in FCN prediction maps. Schock et al. 2020 presented a CNN architecture for 3D shape regression which fed with initial segmentation of knee bone and cartilage. The 3D shape model regression was used to predict the PCA parameters and the global transformation parameters that generated the landmark locations. A local segmentation network was used for segmenting the cartilage to produce the final results.

This chapter proposes a multi-task CNN to perform a binary segmentation map of the implant, detect the implant tip point and regress SSM parameters to compute the shape of the implant and infer the Gruen landmarks. The SSM is employed to build a landmark-based shape model from a training dataset and fit this model to a new image using the shape coefficients and pose parameters. The advantages of SSM are combined for both imposing shape constraints and describing the important landmarks in the implant. In addition, the benefits of CNN to extract complex features from images are preserved. Compared with category five approaches, the shape parameters of an SSM are regressed which helps to identify the important landmarks in the implant. This enables further computation and extraction of surrounding regions.

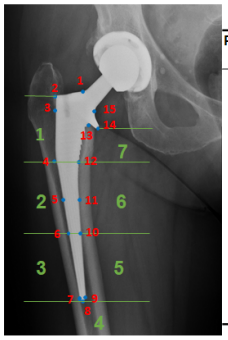
In contrast to other methods, shape prediction is improved by jointly detecting the implant tip point and performing semantic segmentation. In addition, a final alignment of the shape is calculated by applying the ICP algorithm. Compared with category four approaches, the proposed architecture is designed as an encoder-decoder CNN where the features in the encoder part have shape-related information. Therefore, these feature maps are shared by both



Zone 1 : Top of the stem in bone down to the inferior point of GT
Zone 2 : End of zone 1 to halfway to the tip point.
Zone 3 : End of zone 2 to the tip point
Zone 4 : Below the tip point
Zone 5 : Opposite to zone 3
Zone 6 : Opposite to zone 2
Zone 7 : Opposite to zone 1

Point #	Description
1	Superior edge of shoulder
2	The inferior point of GT
3	Halfway to the tip point
4	Tip point
5	opposite to point 3
6	opposite to point 2
7	Collar mid point

(a)



Point #	Description	Point #	Description
1	Neck body interface	8	Tip point
2	Superior edge of shoulder (start of zone 1)	9	Distal end of zone 5
3	Distal edge of shoulder	10	opposite to point 6 (Junction of zone 5 and 6)
4	The inferior point of GT (Junction of zone 1 and 2)	11	Midpoint zone 6
5	Mid point zone 2	12	Opposite to point 4 (Junction of zone 6 and 7)
6	Halfway to the tip point (Junction of zone 2 and 3)	13	Collar inferior point
7	Distal end of zone 3	14	Collar mid point
		15	Collar superior point

(b)

Figure 4.2: (a) Modified definition of Gruen zones. (b) Shape landmarks description.

branches; regression of pose and shape parameters and semantic segmentation. This automates the identification of the Gruen landmarks by constructing the implant shape from the predicted parameters.

The novel contributions of this chapter are: (1) an integrated approach is proposed that allows the segmentation of the implant and automatic detection of the landmarks of interest in the implant. (2) define Gruen zone landmarks and represent the shape of the implant femoral component accordingly. (3) An annotated THR images dataset has been generated to show defined implant landmarks which will be publicly available to enhance the research in this field.

4.2 Anatomical Knowledge

The clinical assessment of the THR postoperative radiographs includes examining the changes in the appearance of implant components and bone. Experienced clinicians depend greatly on their knowledge of the anatomical priors such as the shape and position of the implant and bone for assessing radiograph images. This chapter includes this knowledge in a DL model to segment the implant and detect the important landmarks of the femoral component of the implant.

The main focus of this work is on analysing the femoral component of the implant. The most widely used medical system for evaluating the status of the femoral stem is the Gruen system (GRUEN et al. 1979), which divides the femoral component into seven zones in AP radiograph as displayed in Figure 4.1. Shape landmarks representing the shape of the femoral component and the Gruen zones are introduced. Figure 4.2 (a) shows the definition of Gruen landmarks. Extra landmarks were added to accurately represent the shape of femoral components as presented in

Figure 4.2 (b).

4.3 Shape Model

THR radiograph images significantly vary in appearance depending on the condition of the patient and the complications after THR surgery. SSM is a geometric model that describes the object shape and its variations (Cootes et al. 1995). A model is generated from a training image set that is annotated by a human expert and built from the analysis of the shape variations. The interpretation of a new image requires identifying the parameters that best match the model to the image.

An accurate SSM requires correspondence mapping between shape landmarks. These landmarks are defined using the Gruen zones. The localization of these zones simplifies the analysis of the surrounding region of the implant which, consequently, approximates the shape of the implant and localizes the important landmarks. Figure 4.2 (b) shows a comprehensive description of the defined landmarks. Besides Gruen zone landmarks, additional landmarks within each zone were added to accurately represent the shape of the implant.

After defining the shape landmarks, as a set of N connected landmarks $x = (x_1, ...x_N, y_1, ...y_N)$, the SSM can be constructed using the following steps. First, the mean shape was computed by aligning all the training S shapes together using Generalized Procrustes Analysis (GPA). GPA is an iterative method started by selecting a random shape from the training set as a mean shape, and all shapes were aligned to it. Then, the mean shape was re-estimated, and the alignment was repeated. The process ended when the estimated mean shape was equal to the previous one. The resulting aligned shape is defined as:

$$\begin{pmatrix} x^{'} \\ y^{'} \end{pmatrix} = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{4.1}$$

Where (t_x, t_y), θ and s are the pose parameters (translation, rotation and scaling). The average shape can be estimated by:

$$\bar{x} = \frac{1}{S}\sum_{i=1}^{S} x_i^{'} \tag{4.2}$$

Where $x_i^{'}$ denotes the aligned shape vector, S is the number of samples and $i \in \{1, 2, ..S\}$. This process is presented in Fig. 4.3. The S samples of the training set are shown in Figure 4.3(a)

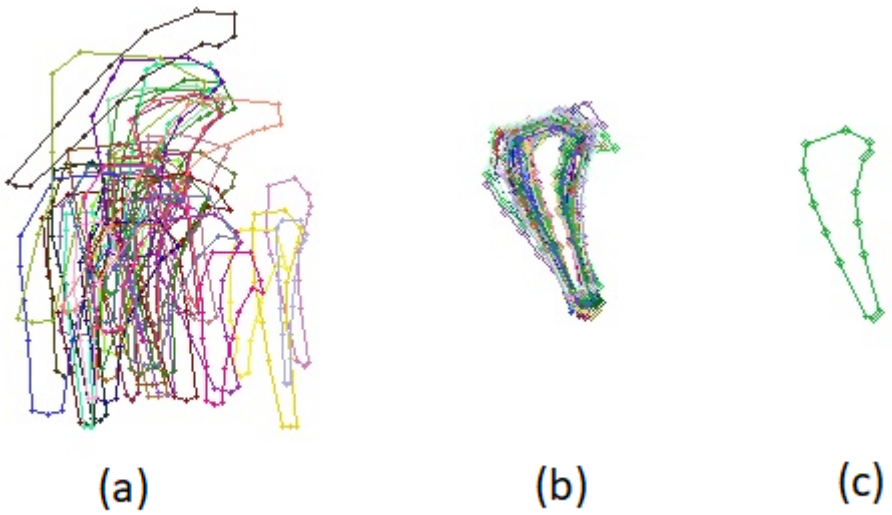


Figure 4.3: GPA steps: (a) Samples of training shapes. (b) Aligned shapes. (c) Mean shape

whereas the aligned shapes x_i are shown in Figure 4.3 (B) and the mean shape $\bar{x}$ is presented in Figure 4.3(C).

Then, the Principal Component Analysis (PCA) was applied to obtain shape variations. Given a set of shape vectors $\{x_i^{'}\}$, the mean shape was computed by using (4.2), and the covariance of the data was computed by:

$$C = \frac{1}{S-1}\sum_{i=1}^{S}(x_i - \bar{x})(x_i - \bar{x})^T \tag{4.3}$$

The eigenvectors $P = \{p_1, p_2, ...p_t\}$ and corresponding eigenvalues λ_t of C represent the directions of variation in the data about the mean. The first M largest eigenvalues were chosen such that:

$$\sum_{i=1}^{M}\lambda_i \geq f_v V_T \tag{4.4}$$

where f_v defines the proportion of the total variation V_T. Assuming that the shape follows a Gaussian probability distribution, the shape can be approximated using:

$$x \approx \bar{x} + Pb \tag{4.5}$$

where P contains the first m eigenvectors and b is a m dimensional vector given by:

$$b = P^T(x - \bar{x}) \tag{4.6}$$

4.4 Dataset Pre-processing

All images were resized to 224×224 px and were normalized by dividing by the largest pixel value (255). Pose parameters (θ and s) and b-coefficients were normalized by min-max feature scaling to values between 0 and 1. The tip point position is used to generate a heat-map image of size 224×224. Each pixel location in the heatmap stores a value representing the probability of that location being a tip point. This is done by creating a Gaussian kernel with $\sigma = 5$ centred at the tip point location.

Online data augmentation was used to increase the dataset size. This was computed by first applying random transformations to the shape parameters as the following: the shape coefficients b were modified by adding a random uniform value $b_{aug} = b + a$ where $a \in [-2, 2]$, random shape rotation $\theta \in [-60, 60]$ and translation by a random value between $[-10, 10]$. Then, the images were transformed according to the computed augmented shape using the Thin Plate Spline Transformation method (Bookstein 1989). The masks and heatmaps were created respectively. In addition, brightness variation $[-0.2, 0.2]$ was applied for augmentation.

4.5 Gruen Net

The proposed Gruen network architecture for detecting Gruen landmarks and performing implant segmentation is presented in Figure 4.4. The input to the network is the X-ray image and it has four outputs: (1) shape parameters b_m. (2) pose parameters θ and scale s. (3) implant tip point (c_x, c_y). (4) segmentation maps. The proposed architecture consists of two branches; the green branch, which is responsible for learning b_m, θ and s, and the yellow branch which learns the binary segmentation map and the tip point heatmap. The grey layers are shared by all tasks. Semantic segmentation and tip point prediction share the same features that are conducted by an encoder-decoder to infer the probability label map. The encoder part includes three residual blocks that consist of two convolution layers with a kernel size of 3×3. Each convolution layer is followed by batch normalization and ReLu activation function. The encoder is followed by the bridge part which consists of one residual block. The decoder part uses both

the features map from the bridge and the skip connections from different encoder blocks to learn the binary classification of each pixel for both segmentation and tip point localization tasks. Finally, the task-specific layers which include convolution and ReLu layers followed by a sigmoid function are added to the network architecture. The regression of the SSM parameters branch starts from the bridge block. It shares three convolution layers and has specific two convolution layers and a Linear layer. Each convolution is followed by batch normalization and ReLu layers.

The network was trained using a weighted sum of multiple loss functions ($L_b, L_{\theta,s}, L_{sh}, L_{c_x,x_y}$ and L_{seg}). These weights were empirically defined. The shape parameters loss (L_b) is defined as the Mean Squared Error (MSE) between the ground truth shape parameters ($b_{i,true}$) and the predicted one ($b_{i,pred}$):

$$L_b = \frac{1}{N}\sum_{i=1}^{N}(b_{i,true} - b_{i,pred})^2 \tag{4.7}$$

The pose parameters loss ($L_{\theta,s}$) are defined as the sum of MSE loss between the ground truth ($\theta_{i,true}, s_{i,true}$) and predicted orientation and scale ($\theta_{i,pred}, s_{i,pred}$):

$$L_{\theta,s} = \frac{1}{N}(\sum_{i=1}^{N}(\theta_{i,true} - \theta_{i,pred})^2 + \sum_{i=1}^{N}(s_{i,true} - s_{i,pred})^2) \tag{4.8}$$

The heatmap regression is employed to detect the tip point. For each image, a heatmap image is formed using a Gaussian filter that is centred at the tip point location. The heatmap loss (L_{hm}) is defined using the Cross-Entropy loss function as:

$$L_{hm} = hm_{true} \cdot \log hm_{pred} + (1 - hm_{true}) \cdot \log(1 - hm_{pred}) \tag{4.9}$$

where hm_{true} is the ground truth label and hm_{pred} is the predicted probability of the point being tip point.

The implant shape is computed using the shape parameters (b_i, θ_i, s_i, and c_x, c_y) as described in section 4.3. The shape loss (L_{sh}) is calculated by the MSE between the predicted shape ($sh_{i,pred}$)) and the ground truth shape ($sh_{i,true}$)):

$$L_{sh} = 1/N\sum_{i=1}^{N}(sh_{i,true} - sh_{i,pred})^2 \tag{4.10}$$

The binary segmentation loss is defined using the Cross-Entropy loss function:

$$L_{seg} = y_{true} \cdot \log y_{pred} + (1 - y_{true}) \cdot \log(1 - y_{pred}) \tag{4.11}$$

where y_{true} is the ground truth label and y_{pred} is the predicted probability.

4.6 Experimental Settings

Multiple experiments have been carried out to validate the proposed method and the effect of each parameter. In addition, different loss functions and hyper-parameters have been explored to obtain the best results. For simplicity, BSM is referred to as the segmentation resulting from the binary segmentation map and SP is referred to as the segmentation constructed from the prediction of shape coefficients, pose parameters and tip points detection.

The purpose of the first experiment is to assess the performance of each task separately; (1) The prediction of a BSM of the image for semantic segmentation. (2) the prediction of the SP for segmentation and landmarks localization. The BSM was achieved by training the main branch of the proposed model (Figure 4.4 the grey and yellow parts for segmentation task only), while the SP predictions were learned by training the grey and green part and yellow part for heatmap prediction. The effect of the data augmentation was also investigated on both tasks.

The second experiment is to evaluate the performance when combining both semantic segmentation and shape and pose parameters prediction in the learning process. This experiment studies the effect of adding shape loss (L_{sh}) that is computed from the shape and pose parameters. The last experiment will study the impact of employing the ICP method to align the segmentation map with the predicted landmarks.

4.6.1 Dataset

To increase the variability of X-ray images, two different hip implant datasets were utilized to construct, train and validate the proposed method: Orthonet dataset (Patel et al. 2021) and in-house dataset. Orthonet dataset is a publicly available dataset that was originally collected for the classification of the implant model type in knee and hip arthroplasty. It consisted of 1191 unilateral hip X-ray images with 8 different models of implant. Part of this dataset (198 images) was intended for implant segmentation. So, it includes the original X-ray images and

Dataset	Type	#images	Example
Orthonet	Depuy-Synthes Corail (with collar)	29	
	Depuy-Synthes Corail (no collar)	30	
	JRI Orthopaedics Furlong Evolution (with collar)	29	
	JRI Orthopaedics Furlong Evolution (no collar)	27	
	Smith & Nephew Anthology	30	
	Smith & Nephew Polarstem (no collar)	29	
	Stryker Accolade II	30	
	Stryker Exeter	30	
In-house	Fracture Type B1	30	
	Fracture Type B2	30	
	Fracture Type B3	29	

Table 4.1: Distribution of the dataset.

the implant mask images. The images have various sizes and all the images represent the normal status of the implant. More details about this dataset and the generation of the implant masks can be found in (Patel et al. 2021). The in-house dataset was generated for automated PFF diagnosis. It consisted of X-ray images after the THR, which is considered normal cases, and X-ray images with various types of fractures. More details about this dataset can be found in 3.2.3.

Due to difficulties of manual annotation of the ground truth, this work has included part of both dataset. A total number of 330 images were used for training and validation of the proposed method. From Orthonet data, approximately, 30 images were randomly selected from each implant model. The remaining images were selected from the in-house data. The choice of in-house images was based on the fracture type. The fracture types B1, B2 and B3 occur within the implant region. Therefore, the images were randomly selected from these types (approximately 30 images per type). Table 4.1 demonstrates the distribution of the dataset.

Ground truth segmentations of implant femoral component and the SSM landmarks were annotated by a clinical expert using the Microsoft VOTT tool. The landmarks were annotated as described in Figure 4.2 (b). Landmarks (2, 4, 6, 8, 10, 12, 14) are the Gruen zone landmarks, while the other points are added to define the implant boundary precisely. The implant masks were generated by filling the area of the defined shape.

4.6.2 Implementation Details

The femoral stem is represented by $N = 15$ landmarks and (θ, t_x, t_y, s) are computed as explained in Section 4.3. The shape model has $M = 15$ modes of shape variation which explains 98% of shape variation. Figure 4.5 shows examples of the shape variations related to the first 15 eigenmodes of the implant.

The dataset was divided into two parts: training and validation, with the ratio 75% : 25%, respectively. Different augmentation methods have been applied to the dataset to minimize the effect of the small dataset size. Refer to section 4.4 for more details.

The network was trained on a Windows machine equipped with 8 GB RAM, Intel(R) Core(TM) CPU @ 3.00 GHz and GeForce RTX 2080 graphics card. It is trained over 200 epoch with AdamW optimizer, learning rate 1×10^{-4}, weight decay 5^{-4} and batch size = 8.

4.6.3 Evaluation Settings

Multiple evaluation metrics were used to validate the proposed method. As explained earlier, the accuracy of the femoral stem segmentation for both outcomes; BSM and SP are evaluated. Dice coefficient and Hausdorff distance (HD) were used to evaluate the segmentation results.

Additionally, the performance of the predicted shape coefficients, pose parameters and tip point prediction were evaluated. For the pose parameters evaluation, the absolute error was utilized where the orientation error is defined as $\delta\theta = |\theta_{pred} - \theta_{true}|$ and the scale error is defined as $\delta s = |s_{pred} - s_{true}|$. The Euclidean distance was used to validate the tip point prediction. In addition, the impact of each parameter on the construction of the shape landmarks is analysed by taking into account the ground truth of all parameters except the studied one.

The shape landmarks were assessed using the Normalized Root Mean Square Error (NRMSE). NRMSE measures the average distance between the predicted and the ground truth landmarks normalized by the distance between two adjacent ground truth landmarks (x_{i-1}, x_{i+1})

$$NRMSE = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{(x_i^p - x_i^t)^2 + (y_i^p - y_i^t)^2}}{\sqrt{(x_{i-1}^t - x_{i+1}^t)^2 + (y_{i-1}^t - y_{i+1}^t)^2}} \tag{4.12}$$

Where N is the number of the landmarks, (x_i^p, y_i^p) is the predicted landmark and (x_i^t, y_i^t) is the corresponding ground truth landmark. Furthermore, the Cumulative Error Distribution (CED) was utilized to assess the detection of the landmarks. CED plots the cumulative NRMSE against the proportion of images with an NRMSE of less than or equal to a particular value.

The performance of using augmentation, adding shape loss and applying the ICP algorithm was validated by Dice coefficient, Hausdorff and NRMSE.

4.7 Results

This chapter integrated implant shape into a DL model to segment the implant and detect the Gruen landmarks. To demonstrate the effectiveness of this method, ablation experiments on the THR dataset were performed.

The results in Table 4.2 presented the validity of the proposed method. The upper rows in the table showed the segmentation results computed from the BSM component, while the bottom rows showed the segmentation results computed from the predicted shape and pose parameters

Experiment	Dice (%)	HD (px)
U-Net	74 ± 13.3	16 ± 23.4
BSM	78 ± 23.3	20 ± 23.6
BSM + A	79 ± 24	12 ± 17.9
BSM + SP + A	77.7 ± 22.7	10 ± 14
SBSM + A + $\mathbf{L}_{sh}$	**80 ± 22**	**8.8 ± 10.7**
$\mathbf{SP}_{TXY}$	22.04 ± 24.11	34.7 ± 13
$\mathbf{SP}_{HM}$	56.7 ± 16.8	26 ± 30.6
SP + A	57.4 ± 17.7	24.8 ± 23.4
SP + BSM + A	62 ± 15.7	20 ± 15.7
SP + BSM + A + ICP	66.7 ± 17.7	17.5 ± 17.6
SP + BSM + A + $\mathbf{L}_{sh}$	62 ± 15.2	20.4 ± 9
SP + BSM + A + $\mathbf{L}_{sh}$ + ICP	**69.7 ± 16.7**	**16.8 ± 9.8**

Table 4.2: Dice and HD results for Segmentation computed from BSM (Upper row) and segmentation computed from the constructed implant shape SP (Bottom row) in the ablation studies. The best results are highlighted.

SP. For simplicity, **A** represents data augmentation, $\mathbf{L}_{sh}$ represents the shape loss.

For the BSM, the proposed model provided better segmentation results compared to U-net with a dice score of 78%. The performance was further improved when introducing the data augmentation with a dice score of 79% and HD of 12 px. The segmentation did not improve when joining the shape parameters prediction component in the training, however, introducing L_{sh} resulted in the best segmentation performance with a dice score of 80% and HD of 8.8 px. Figure 4.6 illustrated examples of the binary segmentation results compared to the ground truth segmentation in different experimental settings. Also, the dice score was presented for each image. The predicted segmentation appeared disconnected when utilizing the BSM only, while the shape tends to be connected when joining the regression of the shape parameters, specifically when adding L_{sh}.

The bottom rows of Table 4.2 demonstrated the segmentation results computed from SP task. Two experiments were carried out to compute the implant shape. The first experiment regressed the translation, rotation, scale and shape parameters to compute the implant shape. For simplicity, this experiment is denoted as $\mathbf{SP}_{TXY}$. The second experiment differs from the first one by the computation of the translation parameter which was computed based on the position of the implant tip point. The tip point was predicted using the heatmap regression. This experiment is denoted as $\mathbf{SP}_{HM}$. The regression of shape and pose parameters only that included the regression of the translation parameters (SP_{TXY}) produced poor segmentation results with a dice score of 22%. The performance was enhanced significantly, by approximately

34%, when utilizing the tip point to calculate the translation parameter (SP_{HM}). The performance is further improved by adding data augmentation, with a dice score = 57.4%. When joining the BSM, the segmentation performance was improved in both metrics, Dice = 62% and HD = 20 px. On the other hand, the results have not changed when introducing the shape loss. Applying the ICP algorithm to align the predicted shape to the BSM results produced better shape segmentation with dice = 69.7% and HD = 16.8 px. Figure 4.7 showed some examples in different experiments for the segmentation using the predicted shape. Additionally, a dice score is reported for each example. The shape results were improved with each change to the training method. Furthermore, it is illustrated in the images that when aligning the shape to the BSM the shape outcome is enhanced.

Table 4.3 listed a further validation of the predicted shape experiments by reporting the pose parameters (θ and s) errors, the implant tip point detection error and the constructed shape landmarks error. Regression of pose and shape parameters (SP_{TXY}) provided the best rotation and scale outcomes with $\Delta\theta = 4.48^{\circ}$ and $\Delta s = 0.13$. The same scale error was produced when the BSM was combined with the training and the shape loss was added. However, the error difference among experiments for both orientation and scale parameters was slightly low, $0.3^{\circ} - 1.6^{\circ}$ and $0.1 - 0.2$ scale value, demonstrating that these parameters did not benefit from combined semantic segmentation to some extent. The translation error was measured using the distance between the predicted implant tip point and the ground truth point. The results demonstrated that the translation parameter has improved significantly with each modification to the training method and provided the best result when the BSM joined the training and the L_{sh} was applied. The regression of translation parameters in the first experiment produced a large error. Introducing the tip point heatmap prediction to compute the translation parameters has improved the results from 88 px to 5.11 px. Similarly, the shape landmarks have improved in each alteration and the best outcome has resulted from the last experiment (SP + BSM + A + L_{sh}), with distance error = 0.55 px. The shape landmarks have been considerably enhanced by aligning the constructed shape to the predicted segmentation, which has reduced the error by 0.22 px.

In addition, the impact of the error in each shape component i.e. translation, rotation, scale and B-coefficient on the final reconstruction of the shape landmarks was studied. To this end, shape landmarks were constructed by fixing the values of all shape parameters to the ground-truth

Experiment	$\theta(^\circ)$	scale	Tip point (px)	Landmarks (px)	ICP (px)
SP_{TXY}	**4.48 ± 3.24**	**0.13 ± 0.09**	88.07 ± 19.73	1.54 ± 0.98	-
SP_{HM}	5.80 ± 4.45	0.16 ± 0.12	5.11 ± 31.31	0.80 ± 1.44	-
SP + A	6.09 ± 4.34	0.14 ± 0.10	3.46 ± 23.09	0.71 ± 1.13	-
SP + BSM + A	4.78 ± 4.32	0.14 ± 0.10	2.17 ± 8.28	0.57 ± 0.42	0.36 ± 0.27
SP + BSM + A + L_{Sh}	5.41 ± 4.23	**0.13 ± 0.10**	**1.29 ± 0.94**	**0.55 ± 0.30**	**0.33 ± 0.20**

Table 4.3: Mean and standard deviation for orientation error, scale error, tip point Euclidean distance, the NRMSE for the shape landmarks and after applying ICP method. The best results are highlighted.

Method	Dice (%)	HD (px)
UNet	74.0 ± 13.3	16.0 ± 23.4
Res-Unet	72.3 ± 12.0	17.5 ± 25.3
UNet ++	70.3 ± 13.1	33.0 ± 42.0
Attention UNet	69.0 ± 16.0	44.2 ± 48.1
R2UNet	48.2 ± 17.7	34.0 ± 25.9
CE-Net	55.5 ± 5.50	135 ± 24.1
U2Net	57.1 ± 9.32	124 ± 20.6
Our method	**80.0 ± 22.0**	**8.80 ± 10.7**

Table 4.4: Quantitative results for implant segmentation on the used dataset. Best results are in bold

values except the parameter under investigation which involved the predicted value. Figure 4.8 presented the NRMSE between the ground truth landmarks and the computed shape. The figure indicated that the landmarks error resulting from the error in the translation parameter has improved significantly in each modification. In addition, the B-coefficient error indicated a slight enhancement to the landmarks error. On the other hand, scale and translation errors have a major impact on the landmark error compared to the other parameters.

Furthermore, the performance of landmarks detection is summarised using the CED curves. It can be seen in Figure 4.9 that in both experiments i.e. using the simultaneous training method with data augmentation and by adding shape loss the localization of the landmarks, 80% of the images were below 0.5 NRMSE. On the other hand, $\sim$ 40% of images were below 0.5 NRMSE using the prediction of shape parameters only (SP_{HM}+A and SP_{HM}+A +L_{Sh}). In addition, the maximum error produced by the simultaneous training method was lower than in other experiments.

4.7.1 Experimental Comparison on Test Dataset

To validate the advantages of the proposed method, state-of-the-art networks for both medical image segmentation and landmarks detection were considered as a comparison strategy.

Method	NMRSE (px)
UNet	3.21 ± 1.02
VGG16	3.06 ± 1.35
DenseNet121	2.78 ± 1.07
ResNet50	2.90 ± 1.14
SwinNet	3.00 ± 1.28
The proposed method	**0.33 ± 0.20**

Table 4.5: Quantitative results for Gruen landmarks detection on the used dataset. The best results are in bold

Seven state-of-the-art networks were utilized to compare the segmentation results: U-Net (Ronneberger et al. 2015), Res-Unet (Zhengxin Zhang et al. 2018), U-Net++, (Z. Zhou et al. 2018), Attention Unet (Oktay et al. 2018), R2Unet (Alom et al. 2018), CE-Net (Gu et al. 2019), U2-net (X. Qin et al. 2020). Table 4.4 listed the results of the implant segmentation using different segmentation networks. It is observed that with a small-size dataset, the complex models might be more prone to overfitting, introduce complexity and cannot generalise from limited training samples.U-Net tends to be a better solution because it has a relatively smaller number of parameters compared to other variants. However, employing shape priors has significantly improved the results to 80% dice score.

Regarding the Gruen landmarks detection, the proposed method was compared with various CNN-based networks: U-Net, ResNet50 (K. He et al. 2016), VGG16 (Simonyan and Zisserman 2015), DenseNet121 (G. Huang et al. 2017) and SwinNet (Z. Liu et al. 2021) for predicting the landmarks as direct regression of the points or as heatmap regression. Table 4.5 listed the NRMSE of each tested model. Among the models compared, DensNet performed the best, achieving a 2.9 px NMRSE. However, the developed model achieved a significantly improved landmark detection with a distance error of only 0.33 px NMRSE.

4.8 Discussions

A recent survey demonstrated that combining DL with medical knowledge has a huge impact on the outcomes of several medical image analysis tasks, including segmentation and diagnosis (Xie et al. 2021). Therefore, this strategy is adopted for implant joint images domain aiming to automate the segmentation of the implant and detection of the Gruen landmarks. Despite the challenges imposed by a limited dataset, incorporating implant shape knowledge into the CNN shows precise and valid implant segmentation and Gruen landmarks detection.

In this chapter, the implant shape is defined using the Gruen landmarks definition and a DL method is presented to predict the shape and pose parameters of the implant femoral component and perform semantic segmentation. Compared to typical semantic segmentation where each pixel is binary classified, this approach predicts the shape and pose parameters which link to landmarks representation that can be used in many diagnostic tasks. Diagnosing implant complications depends mainly on the position in relation to the implant. Therefore, the shape landmarks are defined based on Gruen zones, to combine the advantages of both the segmentation and the detection of important landmarks. This is the first algorithm that can detect locations of the important landmarks and segment the femoral component. This has been successfully demonstrated through the comparison of the segmentation and landmarks detection results with the state-of-the-art segmentation models and landmarks detection models. The landmarks localisation results could be considered state-of-the-art results. The dataset used in this work will be publicly available to enhance the research on this domain.

The results of the proposed approach indicated that the regression of the shape and pose parameters is a more challenging process compared to semantic segmentation. The shape and pose parameters regression is performed by training on and predicting a small number of uncorrelated values (19 values) per image using a limited-size dataset, whereas the semantic segmentation is predicted based on a large number of correlated values per image. Replacing the translation parameter regression with translation computed from the prediction of the tip point position has substantially enhanced the shape outcomes. The shape-based data augmentation is used to increase the size of the training dataset. Although orientation prediction did not benefit from the data augmentation, the prediction of the other parameters has improved which impacts positively on the computation of the position of the landmarks.

Combining the training of semantic segmentation with the shape and pose parameters regression has enhanced the outcomes of the segmentation using the constructed shape (see Table4.2). The shared layers between the two tasks enable the learning of more relevant geometric features. It is hypothesised that the introduction of the shape loss will impact the segmentation output of both tasks. It has enhanced the semantic segmentation performance which also makes indirect benefit to the segmentation based on the shape construction by aligning the resulting shape to the semantic segmentation outcome.

This chapter focused on the segmentation and landmark detection of the implant femoral com-

ponents, however, this method can be extended to other implant joints.

4.9 Summary

In this chapter, a new CNN approach is proposed for jointly segmenting the implant femoral component and predicting the Gruen landmarks. These can assist in the analysis of THR radiographs and in the detection and diagnosis of many THR complications including PFF and implant loosening. To this end, the Gruen landmarks were defined and used to represent the shape of the implant femoral component accordingly. SSM is constructed to describe the implant component shape based on the defined landmarks. A new hybrid CNN network has been developed that simultaneously segments the implant component and regresses the pose and shape parameters of SSM. Thus, the implant landmarks' positions are computed from the predicted shape and pose parameters. For this purpose, an annotated THR image dataset has been generated and will be available to enhance the research field.

Several experiments have been carried out to show the effectiveness of the developed approach. These experiments demonstrated that combining semantic segmentation has enhanced the overall outcomes of implant segmentation and the shape landmarks localisation. In terms of segmentation results, the developed approach achieved an overall dice score of 80%, with an HD of 8.8 px. This outperformed the state-of-the-art results achieved using UNet, which had a dice score of 74% and an HD of 16 px. Similarly, for the landmarks localization, the developed approach achieved an NMRSE of 0.33 px surpassing the state-of-the-art results achieved using DenseNet, which had an NMRSE of 2.78 px.

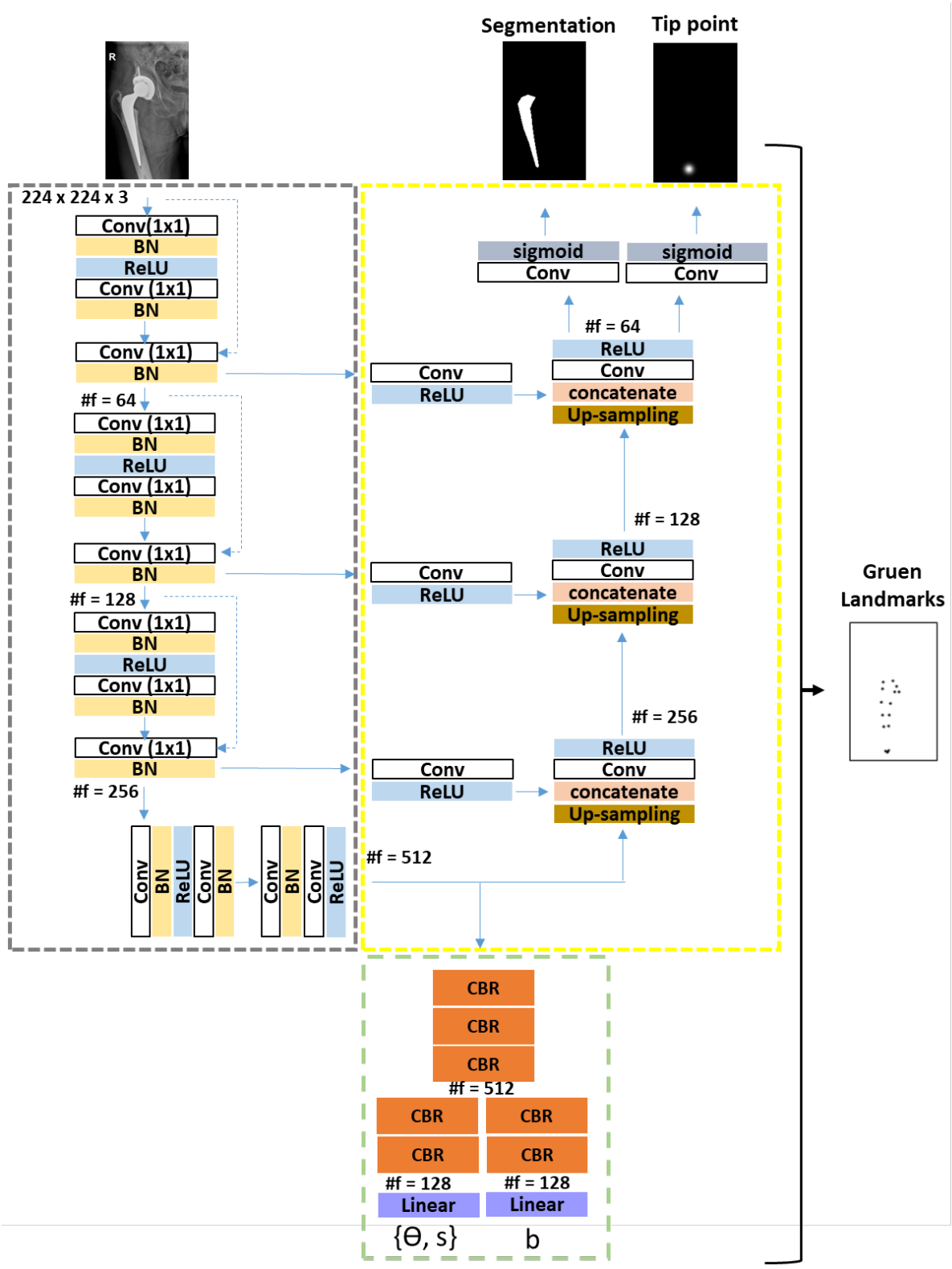


Figure 4.4: The proposed GruenNet architecture has four outputs: shape parameters b, pose parametrs (θ, s), tip point location and segmentation maps. Gray part represents the layers shared by all tasks, the Green part represents shape and pose parameters prediction branch and Yellow represents the decoder part for both binary segmentation maps and tip point detection. CBN refer to Conv, Batch Normalization and ReLU.

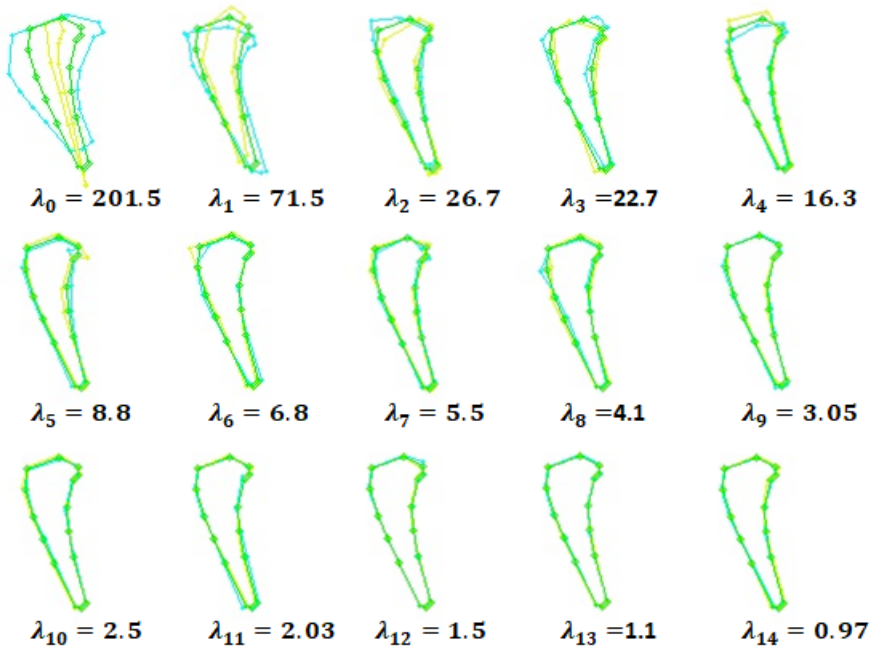


Figure 4.5: 15 modes of shape variations. Green represents the mean model. Yellow represents the deformed shape by $-3\sqrt{\lambda_i}$ and blue represents the deformed shape by $3\sqrt{\lambda_i}$

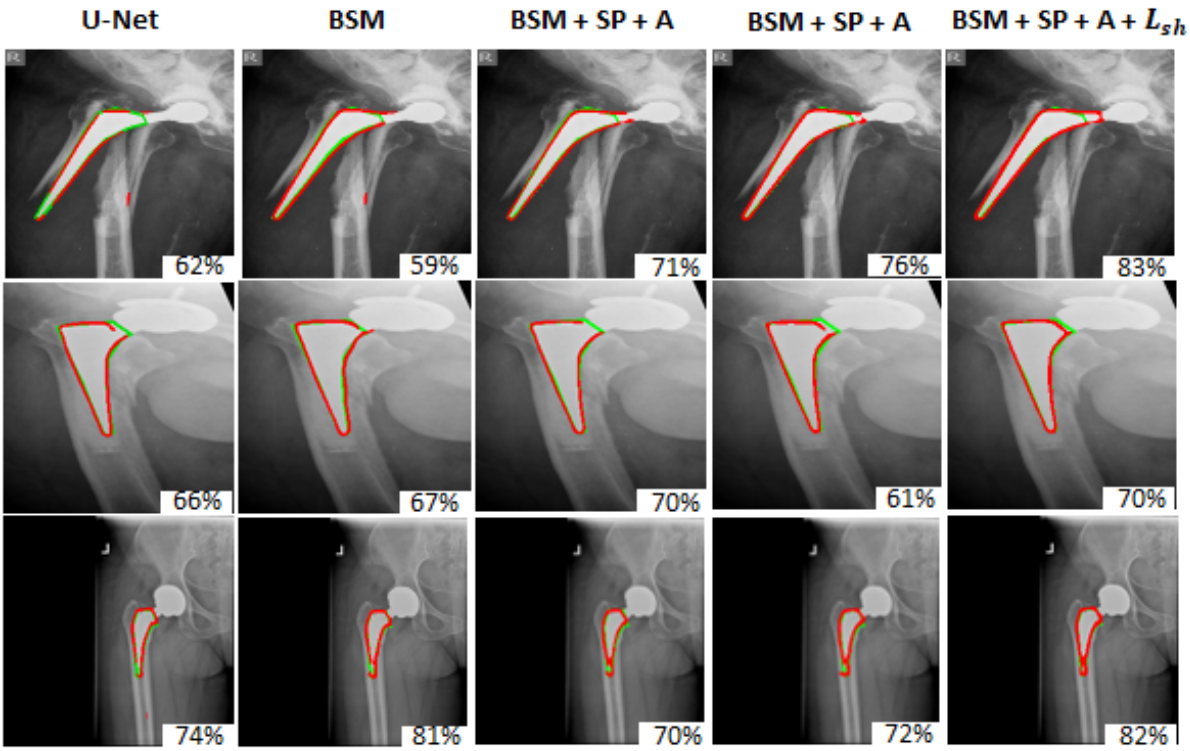


Figure 4.6: Comparison of segmentation computed from BSM in ablation studies. The red is the predicted segmentation and the green is the ground truth. The dice score is presented in each image

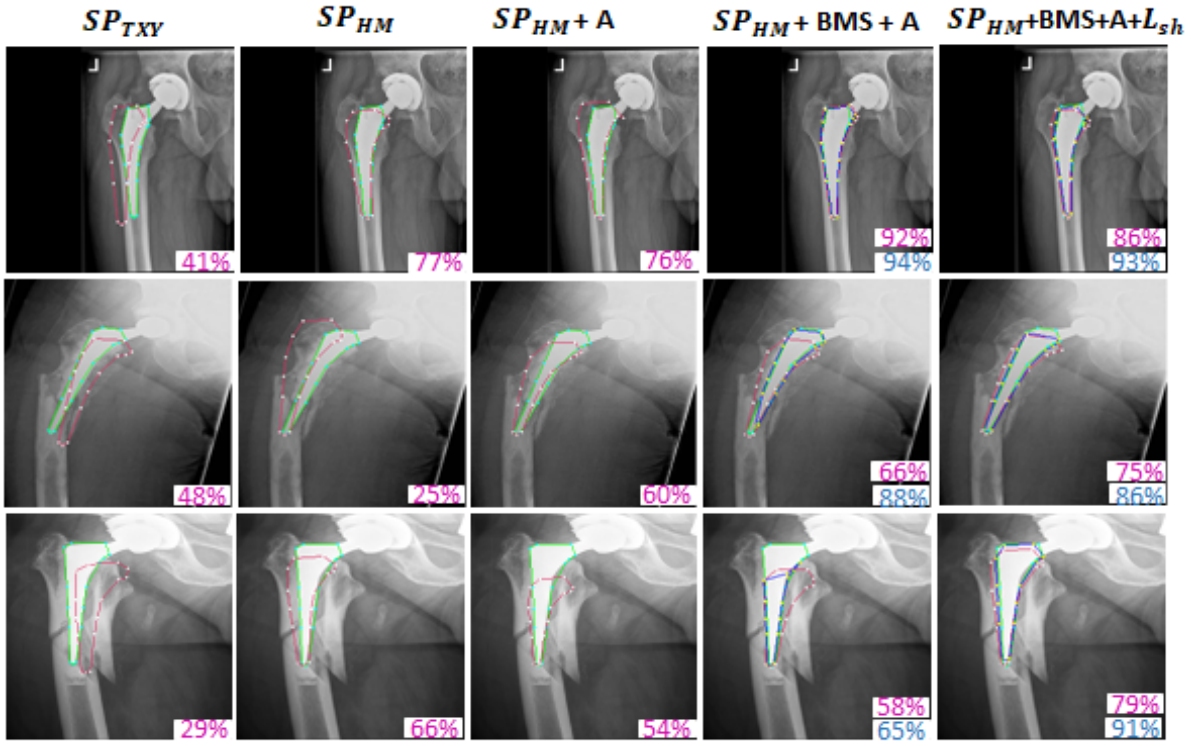


Figure 4.7: Comparison of segmentation computed from SP in ablation studies. The green is the ground truth, the pink is the computed shape and the blue is the shape after applying the ICP algorithm. The dice score is presented in each image.

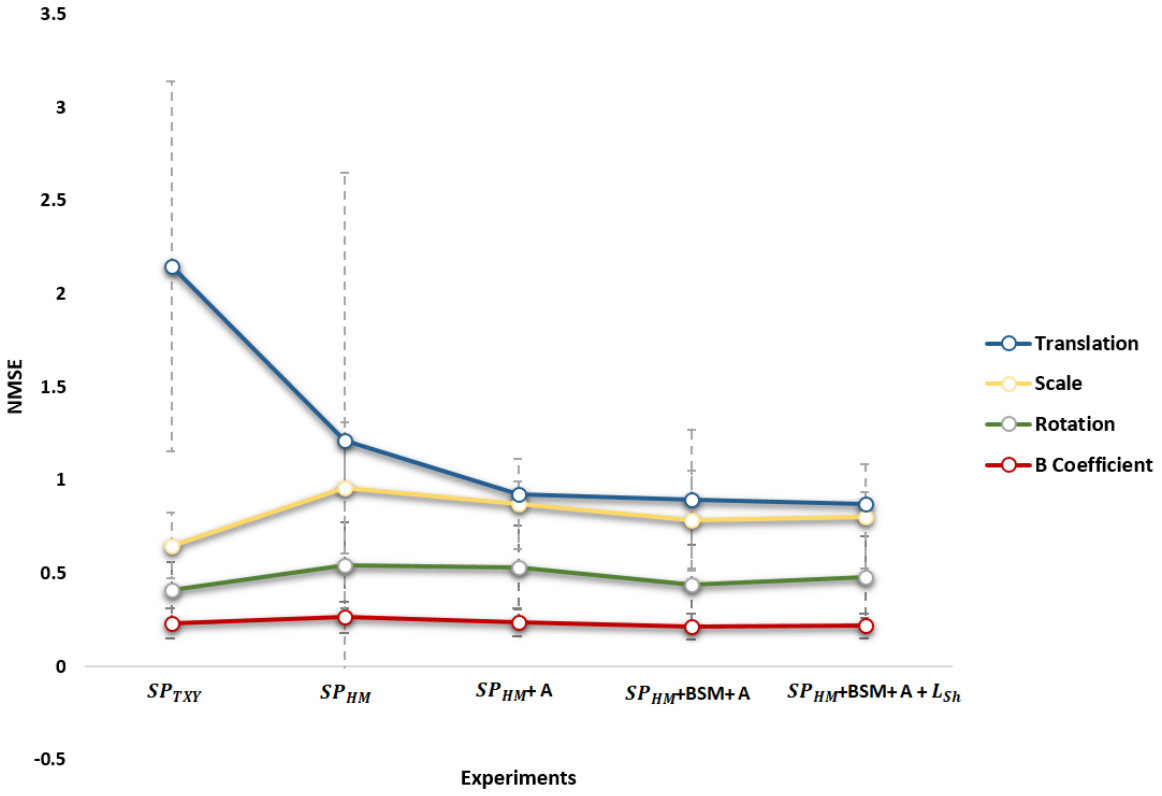


Figure 4.8: The impact of error in translation, rotation, scale and B-coefficient on the computation of the implant shape landmarks. Each plot represents the mean NMSE for the shape computed by fixing all parameters as ground truth values except the studied parameter where the predicted value was used.

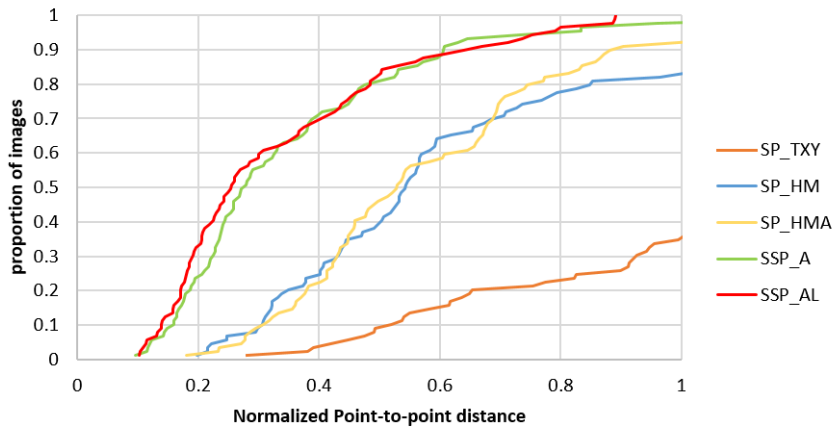


Figure 4.9: CED plot. Comparing the performance of each experiment using point-to-point distance normalized by the two adjacent point distances.

Chapter 5

Incoperation of domain knowledge for Fracture Diagnosis

5.1 Introduction

This chapter aims to develop a CADx tool that automatically identifies PFFs and classifies them based on a clinical classification standard. Besides using such a CADx tool for the classification of the fracture, it can help in efficiently planning the most suitable treatment.

In recent years medical image analysis techniques have significantly improved which has impacted the development of CADx tools for different diseases such as pneumonia detection on chest X-rays (Rajpurkar, Irvin, Zhu, et al. 2017), Alzheimer's disease detection from neuroimaging (Ebrahimighahnavieh et al. 2020) and breast cancer in different image modality (Debelee et al. 2020). The research on bone fractures, when compared to the huge advancements in other medical image domains, is relatively limited. The majority of studies have focused on the detection of the fracture rather than diagnosis i.e. identifying the type of fracture. Several studies have formulated the detection of fracture as a binary classification problem. Earlier work has utilized handcrafted features for the classification such as texture analysis (Chai et al. 2011), a combination of texture and shape features (Umadevi and Geethalakshmi 2012), or digital geometry of the extracted fracture points (Bandyopadhyay et al. 2016). Recently, many studies utilized a typical CNN-based model to detect fractures. For instance, Urakawa et al. 2019, D. H. Kim and MacKinnon 2018 and Rajpurkar, Irvin, Bagul, et al. 2017 trained VGG16, Inception v3 and DenseNet to detect fracture in the proximal femur, wrist and upper extremity,

respectively. Some of these methods used a class activation map to visualize the fracture area (Rajpurkar, Irvin, Bagul, et al. 2017, Varma et al. 2019, V. Gupta et al. 2020).

On the other hand, many studies utilized object detection models to detect and localize the fracture by defining the boundary of the fracture region. Typical object detection models such as Faster R-CNN and modified U-Net have been trained to localize fracture in distal radius X-ray images (Yahalomi et al. 2019) and in wrist X-ray images (Lindsey et al. 2018), respectively. Another work leverages anatomical symmetry cues to build a Siamsea network to detect fractures in pelvic X-ray (H. Chen et al. 2020). First, they applied graph-based landmark detection to identify important landmarks in the pelvic and regressed the line of bilateral symmetry, to flip the image accordingly. The ROI is extracted using specific points. The Siamsea network takes the ROI and flipped ROI as input to produce the fracture probability map.

While some research has been carried out on fracture detection, there have been few investigations into fracture diagnosis. Traditional geometric analysis methods have been investigated for feature extraction and classification. For instance, (Bayram and Çakirotlu 2016) classified diaphyseal femur fractures by first extracting the bone region using the Niblack thresholding method (Niblack 1985). Then, they extracted features, such as area, convex area and angularity, to distinguish between bone and noise segments using SVM. The final classification of the fracture is conducted by feeding multiple features, such as the number of fragments, fractured ends (beginning and end points of fracture region) and fracture region (fracture line), into the SVM classifier. Recently, deep learning techniques have been widely used to develop a model that can learn from data rather than relying on traditional geometric features. Several diagnosis methods trained traditional CNN-based model to classify fracture on image that was manually cropped into a specific bone part such as the proximal femur (Tanzi et al. 2020, Jiménez-Sánchez, Kazi, Albarqouni, S. Kirchhoff, et al. 2018, Lotfy et al. 2019) and proximal humerus (Chung et al. 2018). Other work automated the localization of ROI as pre-processing step using an object detection-based model and then trained the classification model to classify the fracture (Krogue et al. 2019, Jiménez-Sánchez, Kazi, Albarqouni, C. Kirchhoff, et al. 2020). Another work utilized an object detection-based model to localise and classify fractures automatically. S.-J. Yoon et al. 2020 trained a Faster R-CNN model to localize and classify Intertrochanteric femur fracture on the X-ray image. However, these attempts focused on a

particular and small part of the fractured bone e.g. proximal femur by cropping this region manually or automatically, which decreased the visual patterns and improved learning. Unlike these fractures, PFF can appear anywhere along the femur, which increased the visual patterns and made learning the bone structure more complex.

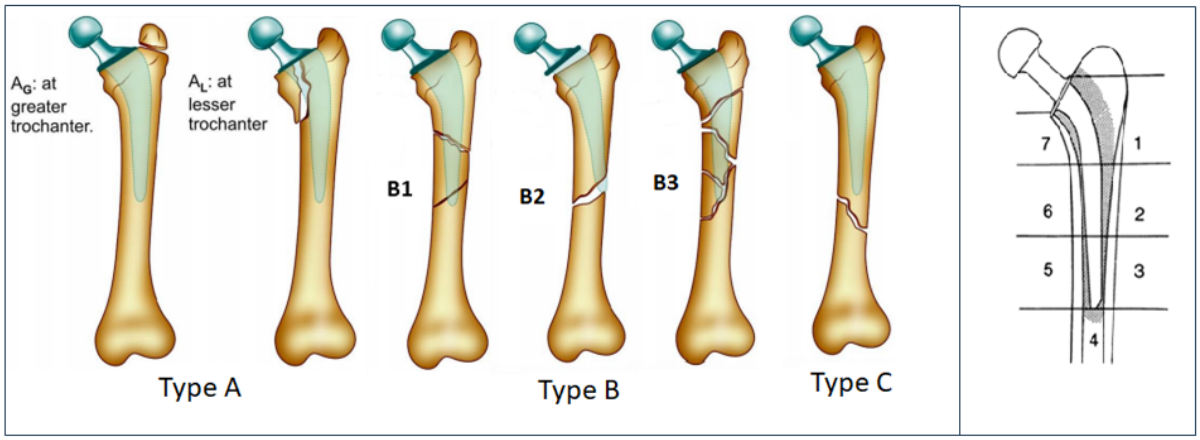
Although typical CNN-based models have shown remarkable performance in various image classification tasks, their effectiveness often relies on a substantial dataset size to fulfilled that advanced stage. The main obstacle toward fracture diagnostic tools is the availability of labelled dataset. Several efforts have been made to solve this problem by utilizing pre-trained models on natural datasets such as ImageNet, and then fine-tune them on specific fracture classification tasks. However, the majority of the published work concentrated on a specific part of the fractured bone e.g. proximal femur or proximal humerus, which reduced the image patterns and led to better learning. Another method used to deal with the lack of data is augmentation techniques where the number of images is increased by creating new images from the original dataset via applying some alterations such as rotation, flips, translations, adding noise etc. Using such techniques might not add new information to the DL model and it could inherit any problem in the original dataset such as bias which is a common issue in medical image data. Recently, adding external information beyond the provided dataset has shown a promising solution to data scarcity issues and increased the performance of the DL model.

In clinical practice, medical doctors made the diagnosis based on their knowledge and experience. The knowledge of medical doctors includes many factors such as the prior knowledge of diagnosis standards and anatomical structure, the way they read the medical image, and the areas they pay more attention to etc. Also, the experience they gained over the years and their practice has impacted the diagnosis accuracy. There is a growing number of medical image diagnosis techniques attempted to add such knowledge into DL model and it showed enhanced performance of diagnosis accuracy such as in Glaucoma Diagnosis (L. Li et al. 2019), thoracic disease diagnosis in chest X-ray (K. Wang et al. 2020) and breast mass classification (X. Li et al. 2020). In fracture diagnosis, there have been few empirical investigations into incorporating medical knowledge into DL. Jiménez-Sánchez, Mateus, et al. 2019 adopted the training patterns of medical students in the training process of the DL model using curriculum learning. As students start with simple tasks and then needed to achieve more complex tasks, the authors assigned a degree of difficulty to each training sample of proximal femur fracture X-ray images.

They demonstrated that when started training the model on 'easy' examples and gradually moving to the 'hard' ones the model can perform better despite having fewer data. Luo et al. n.d. utilized the same approach for the classification of elbow fracture but they proposed a multiview deep curriculum learning rather than a single one. Considering domain knowledge such as the training patterns provide high-level information that is generally used among different diseases. However, low-level information such as features that the clinician commonly analyzed to define the fracture type could be more effective knowledge for incorporating with the DL model to diagnose the fracture. Therefore this chapter includes this information in DL to diagnose PFF and studies the impact of combining such knowledge.

This chapter proposes a knowledge-guided framework that consists of four main feature extraction components and a joint fusion learning component. In clinical practice, PFFs are commonly detected by first examining the femur X-ray images to define the possibility of fracture appearance. Thus, femur X-ray images can be used to train a network to extract the general features of the fracture. It is important to note that the precise identification and localization of discriminative regions greatly enhance region-based feature learning and enable the model to capture and extract relevant information for recognition tasks. Prior medical knowledge can support and enhance this issue. For instance, fracture is usually determined by assessing the bone region surrounding the implant stem. In fact, the fracture's main classification criteria is the anatomical location of the fracture relative to the implant stem. Also, the commonly used protocol to assess the implant and surrounding bone features is to analyse specific zones surrounding the implant known as Gruen zones. Thus, the Gruen landmarks are automatically detected and the zones of interest are cropped. This enables analysis of zone features and fracture features with the knowledge of the location. The proposed framework combines information on pathological features in different regions to provide rich knowledge about the sample. It simulates the visual attention of a clinical specialist and the extensive analysis when making a diagnosis.

The contribution of this work is: (1) the challenge of PFF diagnosis is addressed by developing a novel DL network that leverages clinical knowledge to guide the extraction and fusion of features from the most distinctive regions of femur X-ray images. This approach enables to extraction of comprehensive deep feature knowledge from the images, resulting in a more robust and effective solution. (2) the proposed framework has improved the PFF diagnosis with a higher mean F1 score compared to the state-of-the-art model suggesting its ability to accurately identify



(a) Vancouver classification for PFFs. (b) Gruen Zones

Figure 5.1: (a) fracture classification according to Vancouver system (Schwarzkopf et al. 2013). (b) Femoral component zones according to Banaszkiewicz 2014

positive instances and effectively reduce the false positives or false negatives. (3) define the clinical process of reading the PFF X-ray images and simulated the diagnosis accordingly.

5.2 Medical Domain Knowledge

In fact, identifying the way that clinical experts use for interpret the X-ray images, such as the areas in the image and features of the fracture that they concentrate on, is a challenging task. Therefore, the proposed method uses medical knowledge that was gained from multiple resources:

1. Different medical articles and the commonly utilized standards for categorizing PFFs were studied deeply (GRUEN et al. 1979,ENGH et al. 1990,Ebramzadeh et al. 2003, T. J. McBride and Prakash 2011,Maggs et al. 2021,Powell-Bowns et al. 2021).
2. Observed and participated in the data annotation phase.
3. Discussion of the diagnosis process with multiple clinical specialists.

The commonly used protocol for defining PFF type is the Vancouver classification system. This system defined three main types of fractures (Type A, Type B and Type C) and 6 sub-types (Type AL, Type AG, Type B1, Type B2, Type B3 and Type C) based on three characteristics in the femur X-ray images: (1) the location of the fracture along the implant. (2) stability of the implant. (3) surrounding bone quality. Figure 5.1 (a) illustrates these characteristics. The location of the fracture is an important characteristic which defines the main types of fractures.

Type A includes fractures in the trochanteric area. It involves two sub-categories: Type AG, a fracture within the greater trochanter area, and Type AL, a fracture in the lower trochanter area. Type B includes fractures around the stem. It is divided into three sub-categories: Type B1 when the implant is stable, Type B2 when the implant is loose and Type B3 where the implant is loose and has inadequate bone quality. Type C fractures include any fracture located well below the implant. The common protocol to examine implant loosening is to assess the Gruen zones. GRUEN et al. 1979 evaluated the femoral component interface by defining 7 zones in Antero-Posterior (AP) radiograph of the femoral stem (see Figure 5.1 (b)). The appearance of each zone is evaluated by searching for radio-lucencies lines i.e. gaps between the implant and bone and determining the implant's positional variations in relation to the bone.

From these observations, this section defines the clinical process of reading the PFF X-ray image as the following steps:

1. Browse the femur region: the clinical specialist started by screening the X-ray image to recognize the global visual features and identify the appearance of the fracture.
2. Define the location of the fracture in relation to the implant i.e. in the trochanteric area, around or below the stem.
3. Concentrate on the implant loosening: the clinical specialist focused on each Gruen zone to assess the loosening.
4. Examine the bone stock surrounding the implant.
5. Consider all factors to make a decision.

5.3 Method

The proposed knowledge-guided network of PFF diagnosis consisted of three main steps: (1) Extracted global features from the whole X-ray images. (2) Detected the Gruen zone landmarks and cropped each zone. (3) Constructed two zone branches and trained each one on a different task. Note that all the zones are cropped automatically.

The overview of the proposed method is shown in Figure 5.3. The process of clinical interpretation of PFF X-ray images involved analyzing the global appearance of the femur and implant i.e. whether there is a fracture or not. Then, gradually focused on other factors i.e. location

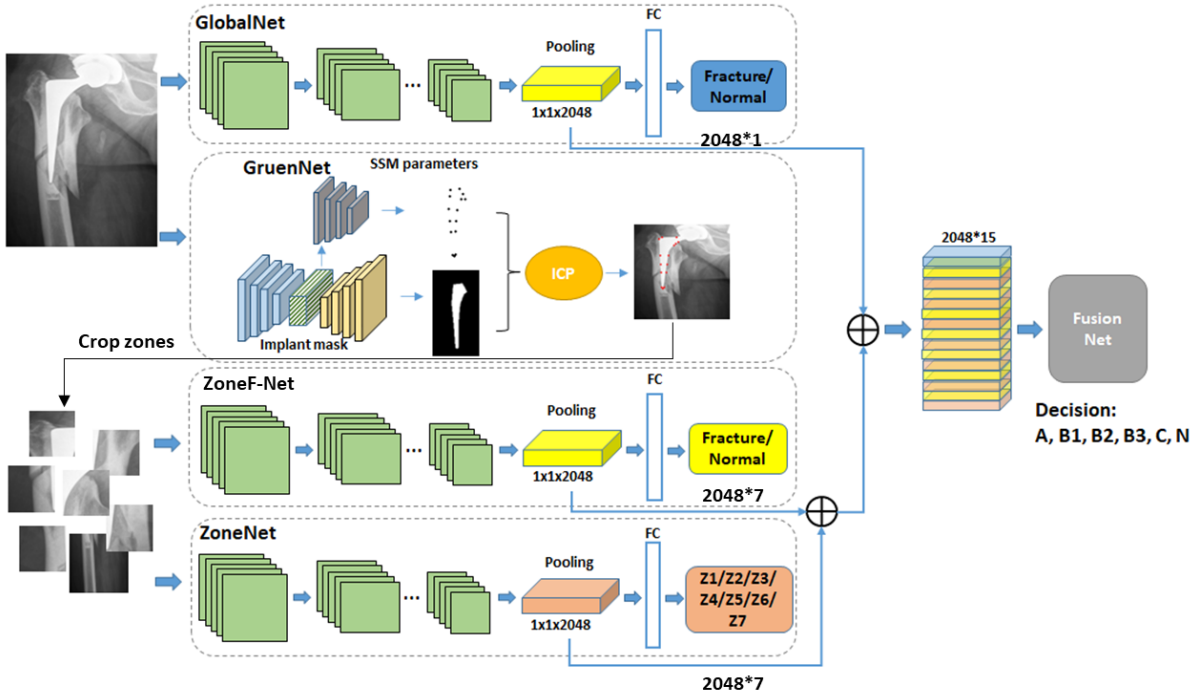


Figure 5.2: The framework of the proposed method. It consisted of three feature branches (GlobalNet, ZoneNet) and landmark detection branch (Gruen Net) and one fusion branch (Fusion classifier). The input to GlobalNet and GruenNet is the full femur image. The input of the ZoneNet is the cropped zones. The input to the fusion classifier is the feature extracted from GlobalNet (F_1), the feature extracted from ZoneFNet F_2 and the feature extracted from ZoneNet F_3.

of the fracture, loosening and bone quality. First, GlobalNet is trained on Femur X-ray images to learn the global visual features of the fracture. Then, guided by prior medical knowledge of the Gruen zones and the important landmarks in the implant, GruenNet is utilized, which was proposed in 4, to detect the Gruen zone landmarks. These landmarks were used to automatically crop the zones of interest (Gruen zones). To concentrate on the other factors i.e. location of the fracture, loosening and bone quality, two sub-models are trained on the extracted zones images. First, ZoneF-Net was trained to analyse the features of the fracture in each zone. Second, Zone-Net was trained to define the zone labels such as Zone1, Zone2 etc. The extracted features from these two sub-models are concatenated which provided knowledge of the location of the fracture and analysed the fracture characteristics. Finally, the fusionNet concatenated the extracted features from the GlobalNet, ZoneF-Net and Zone-net for a final PFF-type classification.

5.3.1 GlobalNet Branch

The GlobalNet branch analysed the global fracture features. The input to the GlobalNet branch was the whole femur X-ray images. A pre-trained ResNet50 (K. He et al. 2016) is utilized as the backbone of this branch. It consisted of 50 learnable layers involving four down-sampling blocks and skip connections between residual blocks. Followed by a global average pooling and Fully Connected (FC) layers for binary classification of the fracture. A sigmoid function was used to normalize the output of FC to $\left[0, 1\right]$ which is defined as:

$$\sigma(x) + Gl = \frac{1}{1 + \exp -x} \tag{5.1}$$

where $\sigma(x)$ is the probability of the input belonging to $\{fracture, normal\}$. In addition, binary cross entropy loss is used which is defined as:

$$L_{Gl} = y \cdot \log \sigma(x) + (1 - y) \cdot \log(1 - \sigma(x)) \tag{5.2}$$

where y is the ground truth label and $\sigma(x)$ is the predicted probability.

5.3.2 GruenNet Branch

GruenNet branch detected the Gruen zone landmarks that were used to extract the zone regions. The input to GruenNet was the whole X-ray image and the output was the implant shape landmarks and the segmentation of the implant. The network architecture was designed as encoder-decoder CNN as shown in Figure 5.2. It consisted of two parts: the part that learned the shape prediction and the one that learned the segmentation maps. In the shape prediction part, the advantage of shape-prior knowledge of hip implant is considered for representing the implant landmarks using SSM. The shape landmarks were defined using the mean shape, the shape and pose parameters $\{b, \theta, s\}$ (refer to section 4.3 for more details). The shape prediction part learned the regression of shape and pose parameters, while the second part learned the binary segmentation map and the tip point heatmap. The two-parts share the encoder. The final step was the alignment of the computed shape which was calculated by applying the ICP algorithm. More details about the network architecture can be found in section 4.5. The shape landmarks consisted of 15 points to represent the implant more precisely (refer to Figure 4.2. The Gruen zones were represented by 7 landmarks (2,4,6,8,10,12,15). For each zone, the top-left and bottom-right corner coordinates of the bounding box are computed and then the region is cropped.

5.3.3 ZoneF-Net Branch

ZoneF-Net branch captured the local features of the fracture in the zone region. Since the clinical specialist analysed the fracture features with the knowledge of the location, the input to ZoneF-Net is the Gruen zones that are located and cropped in the proposed GruenNet. The same CNN structure of the GlobalNet is utilized. The predicted probability is denoted as $\sigma(x)_ZF$ and the loss function as L_{ZF}.

5.3.4 Zone-Net Branch

In fact, the clinical specialist used the anatomical prior knowledge of the zones. Therefore, Zone-Net is proposed to capture the anatomical structure of each zone. The input to this branch was the same as the input to ZoneF-Net i.e. the zones images. The same structure of the GlobalNet was used but with a different FC layer size which was 7 neurons' FC. The predicted probability of the zone name is represented as $\sigma(x)_Z$ and the Cross-Entropy loss as L_Z.

5.3.5 FusionNet

FusionNet branch combined multiple representations from the whole global image to local zone images for the classification of the fracture. Let F_{Gl}, F_{ZF} and F_Z be the features extracted from the average pooling layers in GlobalNet, ZoneF-Net and Zone-Net, respectively. For each zone z_i in image I , F_{ZF_i} and F_{Z_i} are concatenated where $i = 1, 2.., 7$ and F_{Gl}. Then, the concatenated layer was followed by a 6-dimensional FC layer for the final PFF classification.

5.3.6 Training Strategy

In order to enhance PFF classification, the training strategy involved utilizing the pre-trained ResNet-50 network initialization obtained from ImageNet for GlobalNet, ZoneF-Net, and ZoneNet. The proposed GruenNet was used to detect the Gruen landmarks and cropped the Gruen regions which were used as input to Zone-Net and ZoneF-Net for the fine-tuning process. During the fine-tuning process of one branch, the weights of the other branches remained unchanged. In order to leverage the discriminative features presented in the original image and the zone images, the output of the last pooling layer of each network was used as input to the FusionNet. Let P_{Gl}, P_{ZF} and P_Z denoted the average pooling layer in GlobalNet, ZoneF-Net and Zone-Net, respectively. The concatenation of these features was fed to train the FusionNet whereas the weights of the other models were fixed.

5.4 Experimental Settings

Multiple experiments have been carried out to evaluate the proposed method and the effect of incorporating domain knowledge in fracture diagnosis. Various loss functions and hyper-parameters have been investigated to achieve the best outcomes. All experiments were generated using both the ground truth Gruen landmarks and the predicted landmarks i.e. the Gruen zones were calculated and cropped using the ground truth landmarks or the predicted ones. The performance of the classification of the PFF using each classification branch separately which are GlobalNet and ZoneF-Net is assessed. These experiments were denoted as Gl and ZF, respectively. To study the impact of adding more information to the performance of the classification, the components are joined and evaluated gradually. The combination experiments are denoted as follows: $GlZF$ for combining Global-Net and ZoneF-Net branches, ZFZ for combining Zone-Net and ZoneF-Net branches and $GlZFZ$ for combining Global-Net, ZoneF-

Net and Zone-Net branches.

5.4.1 Dataset

An in-house hip implant dataset was used to construct, train and validate the proposed method. More details about the dataset can be found in 3.2.3. 389 images were used for the training and validation of the suggested approach. The dataset consisted of 48 images Type A, 80 images Type B1, 87 images Type B2, 37 images Type B3, 67 images Type C and 70 images Type N. The X-ray images were of various sizes, orientations and implant types. The images included either a partial region of the femur or the full femur with the appearance of a partial region of the pelvic.

The ground truth was generated for each branch depending on each branch's goal task. For the fracture classification, two clinical experts contributed to image annotations and provided class labels. The annotation process was explained earlier in 3.2.3. Considering the detection of the Gruen landmarks i.e. GruenNet, ground truth masks of implant femoral component and the SSM landmarks were annotated by a clinical expert using the Microsoft VOTT tool. The implant shape was represented by 15 landmarks, 7 of these landmarks are the Gruen zones landmarks which are (2, 4, 6, 8, 10, 12, 14). Further details of the annotation process were explained in 4.6.1.

For both ZoneF-Net and Zone-Net, the input images were the cropped Gruen zones. The images were generated based on the Gruen landmarks. The top-left and bottom-right corner positions of the bounding box are computed and cropped the region accordingly. It is observed that few cases of PFF X-ray images contained part of the femur i.e. the distal part is not included in the image which represented zone 4. Thus, the missing image i.e. zone 4 is computed by selecting several zone 4 images from the dataset. This estimates the missing value in some cases.

For ZoneF-Net, the ground truth label for each image was Fracture or Normal. The zone was considered a Fracture when it consisted of one or more fracture lines, otherwise the image was considered Normal. The number of images used in this task was 221 whole images which generated 1547 zone images, 780 of type fracture and 767 of type Normal. Random samples are excluded in order to balance the distribution of classes. For Zone-Net the whole dataset was used which generated 2723 images. The images were labelled by the zone name (Zone1, Zone2, Zone3, Zone4, Zone5, Zone6, Zone7). The number of images per zone was 389 images. Note

that the zone images and labels were generated automatically.

For each task, the dataset was divided into two parts: training and validation, with the ratio 75% : 25%, respectively. The validation set is used as a testing set.

5.4.2 Implementation Details

All the models were trained on a Windows machine equipped with 8 GB RAM, Intel(R) Core(TM) CPU @ 3.00 GHz and GeForce RTX 2080 graphics card and were implemented on Pytorch. GlobalNet was trained on X-ray images down-sampled from the original size to $224 \times 224px$. The classes consisted of normal and fracture. Multiple data augmentation techniques were applied including flipping, rotation and scaling. The architecture details of GruenNet can be found in section 4.6.2. Both ZoneF-Net and Zone-Net were trained on images resulting from GruenNet and were resized to $224 \times 224px$. Data augmentation techniques, such as rotation and scaling were used. The classes in ZoneF-Net included Fracture and normal, while classes in ZoneNet consisted of the zones names (Zone1..., Zone7). For optimization, SGD is used. All the models were trained until convergence (100 epochs). The batch size was 8, momentum 0.9 and the learning rate was set to 1×10^{-2}.

5.4.3 Evaluation Settings

The performance of the developed method is assessed using multiple confusion metrics, applying them individually to each class. The reported metrics represent the average performance across all classes, providing a comprehensive evaluation of the model's performance across different categories. The accuracy of the proposed model was computed by dividing correct predictions by the total number of samples in the dataset. It can be defined as:

$$Accuracy = \frac{TP_1 + TP_2 + ...TP_k}{N} \tag{5.3}$$

where TP_k is the number of true positives in class k and N is the total number of samples. Also, precision was used which measured how many of the predicted positives were actually positive. It can be defined as :

$$Precision = \frac{TP_k}{TP_k + FP_k} \tag{5.4}$$

where TP_k is the number of true positives in class k and FP_k is the number of false positives in class k. In addition, the Recall metric was computed to measure the percentage of correctly predicted positives. It is defined as:

$$Recall = \frac{TP_k}{TP_k + FN_k} \tag{5.5}$$

where FN_k is the number of false negatives in class k. F1-score was also used to summarise the overall performance of the classifier by computing the harmonic mean of precision and recall. It is defined as:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precisison + Recall} \tag{5.6}$$

Additionally, the Specificity metric was applied to measure out of all the predicted negatives how many are actually negatives. It is defined as:

$$Specificity = \frac{TN_k}{TN_k + FP_k} \tag{5.7}$$

Where TN_k is the number of true negatives in class k.

Area Under the ROC curve (AUC) was also used to evaluate the proposed methods. It represented the probability that the model ranked a randomly chosen positive sample more highly than a randomly chosen negative sample. The AUC ranges from 0 to 1, where an AUC of 1 indicated a perfect classifier. In addition, the Precision-Recall curve and AP for each class were computed. The Precision-Recall curve showed the trade-off between precision and recall for different classification thresholds, while AP is the area under the Precision-Recall curve. AP ranges from 0 to 1, where an AP of 1 indicated a perfect classifier.

5.5 Results

To show the efficiency of the proposed method, different ablation experiments were performed on the THR dataset. In addition, a comparison to the state-of-the-art methods is presented in this section.

In order to select the best backbone for GlobalNet, ZoneF-Net and Zone-Net, different pre-trained models have been utilized as a backbone in the proposed architecture. The considered

Backbone	VGG16							DenseNet121							ResNet50							ViT-32						
Classes	**A**	**B1**	**B2**	**B3**	**C**	**N**	**avg**	**A**	**B1**	**B2**	**B3**	**C**	**N**	**avg**	**A**	**B1**	**B2**	**B3**	**C**	**N**	**avg**	**A**	**B1**	**B2**	**B3**	**C**	**N**	**avg**
Precision	86	31	43	44	48	94	57.7	83	45	46	33	78	100	64.2	70	50	42	50	86	100	66	100	46	54	38	71	94	**67**
Recall	50	25	48	44	69	94	55	42	50	62	33	63	100	53.3	58	40	52	78	75	88	**65**	25	65	67	33	63	88	56.8
F1-score	63	28	45	44	56	94	55	56	48	53	33	69	100	59.3	64	44	47	61	80	94	**65**	40	54	60	35	67	91	57.8

Table 5.1: Comparison of utilizing different pre-trained CNN models as a backbone of the proposed architecture. The best results are highlighted.

Methon	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score(%)
ResNet50-W	94	87	96	96	90.5
ResNet50-Z	78	78	78.5	78.5	78.5
ResNet50-ZN	98.7	96	95.6	99.4	95.4

Table 5.2: **ResNet50-W:** the results of fracture identification using the original X-ray images. **ResNet50-Z:** the results of fracture identification in each zone image. **ResNet50-ZN:** The results of the classification of the zone (zone1, zone2. . . zone7).

models were VGG16, DenseNet121, ResNet50 and ViT-32. The aim of the proposed architecture is to improve the discrimination of each fracture type. Therefore, the performance of each model was evaluated using the Precision, Recall, and F1-score metrics. Table 5.1 reported the results of utilizing different backbone models. VGG16 and DenseNet121 showed lower precision compared to ResNet50 and ViT-32, with ViT-32 outperforming ResNet50 by a slight margin, achieving a precision score of 67%. However, ResNet50 achieved higher average recall and F1-score of 65% for both metrics, indicating its ability to distinguish positive samples correctly and provide a greater harmony between recall and precision. Therefore, ResNet50 was selected as a backbone for the proposed framework.

To analyse the performance of each branch, i.e., the feature extraction sub-networks GlobalNet, ZoneF-Net and Zone-Net were examined. Table 5.2 presented the results of each network. For simplicity, ResNet50-W referred to binary classification for the whole X-ray images in Global-Net, ResNet50-ZF referred to binary classification for the cropped zone images in ZoneF-Net and ResNet-ZN referred to the classification of the zone image to which zone it belonged, i.e., {Zone1,...,Zone7}. The table demonstrated that detecting fractures in the original image provided an accuracy of 94%, outperforming the detection of fracture in each region by approximately 16% in overall performance. This might affect the quality of the features extracted in ZoneF-Net. The zone classification model (ResNet50-ZN) demonstrated a high accuracy of 98.7% and achieved balanced performance with a precision and recall of $\sim$ 96% and F1 score of 95.4%, highlighting its effectiveness in accurately classifying the zone while minimizing false positives and false negatives.

To assess the effectiveness of the proposed framework, Table 5.3 showed the outcomes of each

Method	Accuracy(%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)
GL (baseline)	85.7	56.7	53.5	91.3	53.5
ZF	87.5	64.2	62.8	92.2	62.7
ZFZN	87.2	64.3	62.5	91.8	62.3
GLZF	87.2	62	61.7	91.8	61.7
GLZFZ (GT)	**88**	**66.3**	**65.2**	**92.3**	**65.2**
GLZFZ (pred)	85.2	58.5	57.3	91.2	57

Table 5.3: The results of the fracture diagnosis using each component separately as well as adding components to the framework in the ablation study. The best results were highlighted.

component, GL and ZF separately, in fracture diagnosis, as well as the impact of combining multiple components; ZFZ, GLZF and GLZFZ. The table also reported the results using the ground truth zones (GT) and using the predicted zones (pred) for demonstrating the impact of zone prediction error. It is apparent from the table that there is a slight difference in the accuracy among the models, with GLZFZ achieving the highest accuracy of 88%, followed closely by ZF, ZFZN and GLZF, which all had similar accuracies ranging from 87.2% to 87.5%, while GL had relatively lower accuracy of 86.7%. Considering the ability to accurately detect positive cases of a particular type of fracture, GL showed a lower recall score of 56.3%. The performance increased by utilizing the zones images with a recall score of $\sim$ 63% for both ZF and ZFZN, while it slightly decreased when combining GL and ZF and reached its highest score when combining all components GLZFZ, with value of 65.2%. Among the evaluated experiments, combining GlobalNet, ZoneF-Net and Zone-Net, GLZFZ, showed the highest performance when using the GT zones images with an accuracy of 88%. It achieved a precision of 66%, a recall of 65%, F1-score of 65%, and a specificity of 92%, highlighting its better ability to classify and distinguish the fracture samples accurately. The results also demonstrated that the fracture diagnosis performance was influenced by the Gruen zones detection error, leading to a decrease of 2.8% in the overall performance. However, it still showed better Precision, Recall and F1-score compared to the baseline model.

In addition, Figure 5.3 showed the impact of zone prediction error in the classification accuracy by presenting the results of each fracture type using different resolutions of the detected zones compared to the GT zones. For all classes except Fracture type B2, the classification results of predicted zones (th1,..., th4) consistently showed lower accuracy compared to GT zones, indicating a decrease in classification performance caused by zone prediction error. Normal images consistently achieved high accuracy across all resolutions, ranging from 92% to 96%. The classification accuracy of Fracture types A, B1, B2 and Normal varied slightly across different

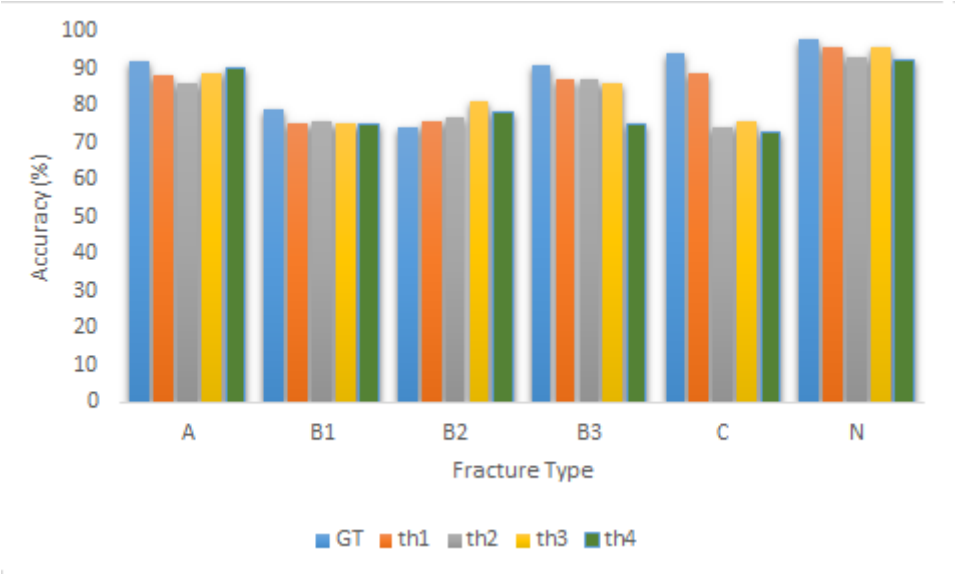


Figure 5.3: The classification accuracy of each fracture type obtained by the proposed method using the zone ground truth (GT) and different resolutions of the predicted zone th1, ..., th4.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	Specificity(%)	AUC(%)
VGG-16	84.7	59	52.3	51.2	90.3	81.5
Densenet121	85.2	56.2	52.7	53.8	90.8	79.8
Resnet50	85.7	56.7	53.5	53.3	91.3	84
ViT-16	80.7	43.5	39.3	40	88	68.5
ViT-32	84.3	51.7	47.7	46.7	90.3	78.5
SwinT	82.7	49.2	46	46.8	89.5	74.5
GLZFZ (GT)	**88**	**66.3**	**65.2**	**65.2**	**92.3**	**89.8**
GLZFZ (pred)	85.2	58.5	57.3	57	91.2	86.3

Table 5.4: Comparison of the classification results between the proposed method and the state-of-the-art classification networks

zone resolutions. In contrast, Fracture types B3 and C showed high variations in performance among different resolutions with the highest classification accuracy obtained by th1 with 87% and 89% for Type B3 and C, respectively.

To validate the advantages of the proposed method, Table 5.4 presented a comprehensive summary of the performance outcomes of various state-of-the-art image classification networks, including ResNet50 (K. He et al. 2016), Densenet121 (G. Huang et al. 2017), VGG16 (Simonyan and Zisserman 2015), Vision Transformer 16 and 32 (Dosovitskiy et al. 2020) and Swin-Transformer (Z. Liu et al. 2021). Accuracy, precision, recall, F1-score, specificity, and AUC for each model were considered as assessment criteria when comparing the proposed framework to these models. The proposed method achieved an accuracy of 88% using GT zones, outperforming all the other methods. Utilising predicted zones, it outperformed VGG-16, ViT-16, ViT-32, and SwinT with 85.2 % and demonstrated comparable accuracy with Resnet50 and

Densenet121. In terms of precision, the proposed method showed a significant improvement compared to other methods with a 66.3% precision score when using GT zones which indicated its ability to better identify positive cases. In addition, utilizing predicted zones a value of 58.5% is achieved, surpassing all models except for VGG-16 with a similar precision of 59%. The recall scores ranged from 39.3%, achieved by ViT-16 to 53.5% achieved by Resnet50, with the proposed method showing a higher recall of 57.3% using predicted zones and 65.2% using GT zones. Similarly, the proposed method reported better F1-score for both using GT and predicted zones with 65.2% and 57%, respectively. The F1-score for the other models varied from 40%, achieved by ViT-16, to 53.8%, achieved by Densenet121. Furthermore, the AUC of the proposed method was 89.8% using the GT zones and 86.3% using the predicted zoned, demonstrating its efficiency in distinguishing between various classes. Notably, ViT-16 presented the lowest AUC among the models, followed by SwinT, ViT-32, Densenet50, VGG-16 and Resnet50, with AUC scores of 68.5%, 74.5%, 78.5%, 79.8%, 81.5% and 84%, respectively.

To further analyse the performance of the proposed framework compared to the state-of-the-art models, Figure 5.4 presented the precision-recall curve and AP for each model, each class separately. Our method was outstanding with the highest precision and recall values for normal case images, as indicated by the precision-recall curve and the AP of 100, establishing its superior capability in accurately identifying Normal images. The other methods also presented a notable performance in identifying Normal cases compared to other classes with an AP ranging from 0.8 using ViT-32 to 0.99 using Densenet121. Among all fracture types, Type C fractures were the most accurately identified by all models, noting that the proposed method outperformed Resnet50, the best-performed model, by an AP score of 9 using predicted zones and an AP score of 14 using GT zones. In fracture type A, the proposed method also showed significant improvement in identifying this type compared to all the other models with AP of 66 and 69 for using GT and predicted zones, respectively. The other models tended to misidentify the images of type A fractures with AP varied from 23, achieved by ViT-16 to 45 achieved by Densenet121. Similarly, for fracture Type B2, the proposed method delivered the best outcomes among the other state-of-the-art models with an AP of 59 using either GT or predicted zones. In terms of fracture types B1 and B3, the precision-recall curve figure showed a gradual decrease in precision indicating the lower ability to identify correct classes, however, the proposed method achieved the best AP of using GT zones for defining class B1.

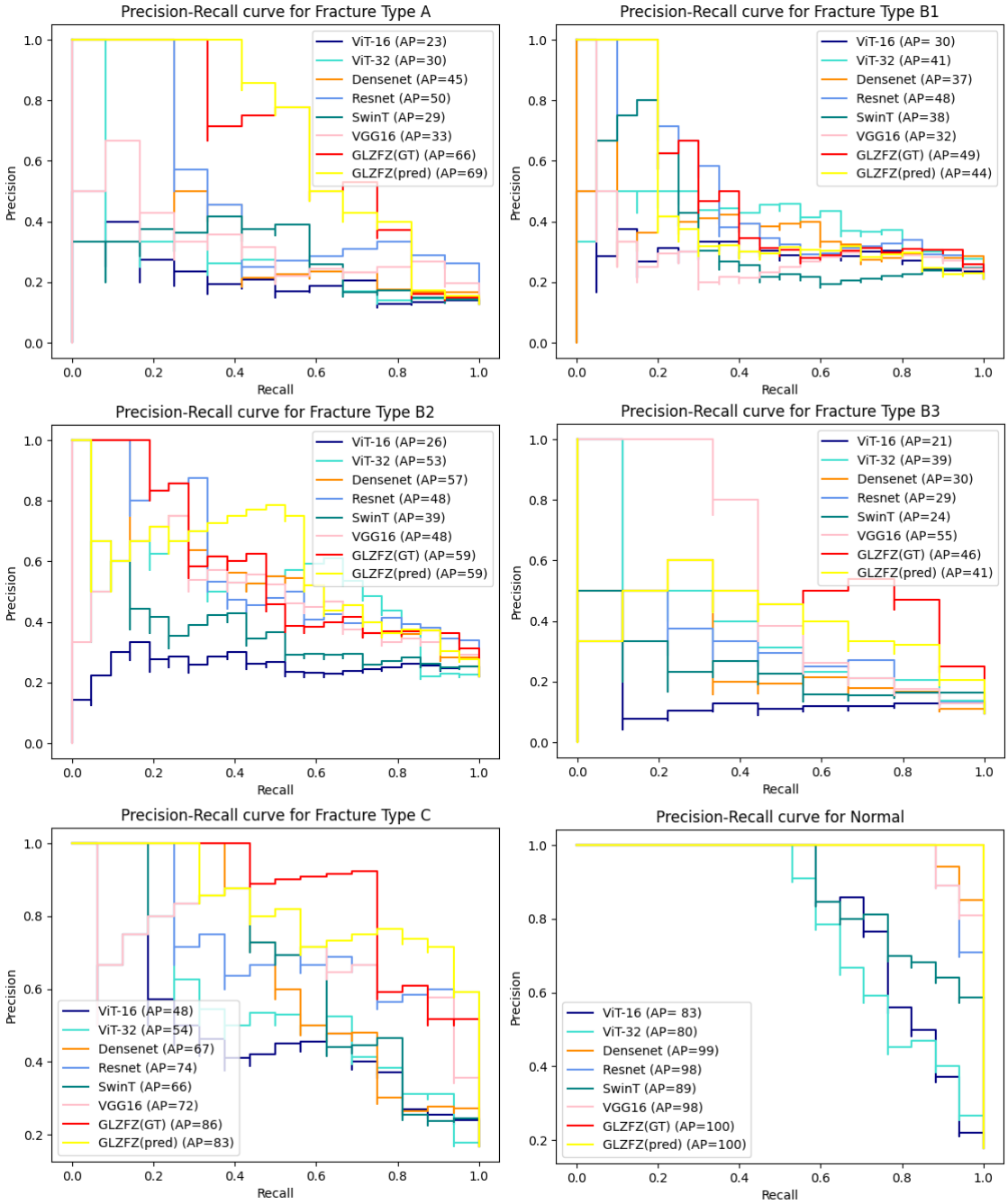


Figure 5.4: Comparision of Precision-Recall curve and AP for each class.

These results showed that the proposed method outperforms various state-of-the-art models and highlighted its potential for effective PFF diagnosis.

5.6 Discussion

PFFs significantly contribute to the need for revision hip arthroplasty, particularly in the elderly population, where an exponential increase in THR rates is observed. The management and treatment planning for PFFs depend mainly on the type of fracture. The diagnosis of PFF depends on the manual assessment of X-ray images by a radiologist. However, 90% of PFF radiology reports omitted important radiographic details, which delays diagnosis (Marshall et al. 2017). This chapter automates the detection of PFFs and the classification of their types which substantially can assist in fractures management and treatment planning. The process of interpreting X-ray images of THR is outlined by incorporating the clinical protocol for defining PFFs types and highlighting the specific regions of interest that clinical specialists usually focus on. This process is simulated and a DL-based method is developed to extract multiple features and fused them to accurately diagnose fractures in THR images.

Despite the challenges of a limited and imbalanced dataset, the proposed framework outperformed several state-of-the-art networks. These findings demonstrated the enhancement in the diagnosis results using this approach and its potential for accurately diagnosing PFFs. In addition, it offered promising prospects for improving the effectiveness and reliability of PFF diagnosis.

Compared to typical classification models that extracted features from the whole X-ray image only, the developed method introduced new information by combining multiple features extracted from both the whole X-ray images and specific regions of interest, the Gruen zones. This addition of features increases the complexity of the method compared to traditional models; however, it leads to a significant improvement in performance. Specifically, when incorporating GT zones, an enhancement of approximately 10% in precision, 12% in recall, and 11% in the F1 score was observed. Similarly, when utilizing predicted zones, improvements of approximately 2% in precision, 4% in recall, and 3% in the F1 score were achieved.

When considering each feature extraction component within the presented framework, the Global-Net and Zone-Net exhibited high performance in binary classification and zone name

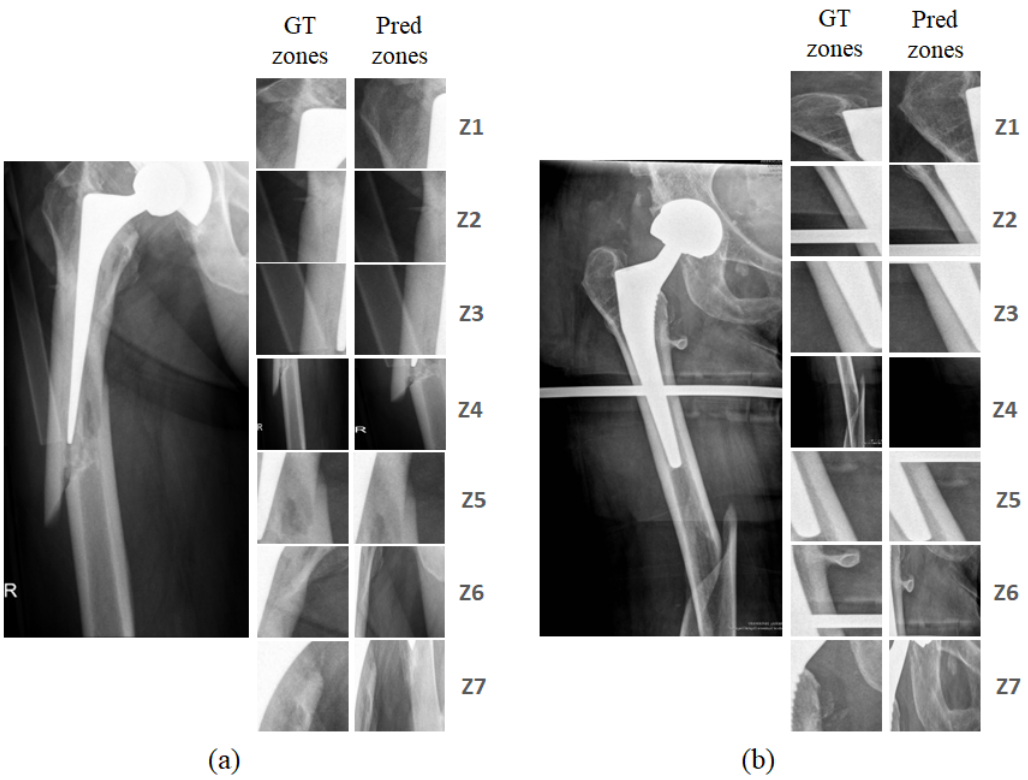


Figure 5.5: Examples of GT zones and predicted zones of each X-ray image.

classification, respectively. However, the ZoneF-net provided lower results in the binary classification of Gruen zones, possibly due to errors in the detected zones. Nevertheless, when combining the features from each zone, ZoneF-net demonstrated better classification performance than the baseline GL. Combining features from Global-Net with ZoneF-net or combining Zone-net with ZoneF-net did not have a significant impact on the results. However, when all components were combined, there was a notable improvement in the classification of PFF.

The localization error in Gruen landmarks can lead to the exclusion of some parts of the bone or fracture region, ultimately impacting the overall classification performance. Figure 5.5 showed examples of the GT zones and the predicted ones for each X-ray image. Even though the predicted zones in Figure 5.5 (a) do not accurately match GT zones, it captured the majority of the fracture properties which led to correct classification results. On the other hand, in the case of the X-ray image presented in Figure 5.5 (b), the incorrect classification may be related to the missing important fracture properties, such as features in zone 4.

The goal of the proposed method is to enhance the detection of the fracture and the discrimination performance for each class. In clinical practice, identifying subtle or hairline fractures poses a challenge. However, our model demonstrates proficiency in this task, achieving an

accuracy of 94%. It is important to note that the testing data consisted of various types of fractures, ranging from clearly visible fractures to tiny hairline fractures. In addition to identifying fractures, discriminating between fracture classes presents a unique set of challenges in clinical scenarios specifically distinguishing between type B fractures i.e. B1, B2 and B3. When considering the model performance of each class individually, it is observed that normal images achieved the highest discrimination results, followed by fracture types C and A. This difference in performance could be attributed to the distinct locations of each type. Fracture type A can appear in Zone 1, 2, or 7, while type C commonly appears in Zone 4. On the other hand, types B1, B2, and B3 can be found in the same zones, namely Zone 2, 3, 5, and 6. Discriminating between B types may present a greater challenge compared to distinguishing between types A and C. However, the developed model provided promising results in distinguishing these types achieving an AP score of 49%, 59% and 46% for classifying fracture type B1, B2 and B3, respectively. While these scores indicate a notable performance, it is essential to contextualize them within the challenges of clinical practice, recognizing the complexity of precisely categorizing fracture types. Continuous refinement and validation, incorporating factors like bone density data, might be valuable for enhancing the model's classification capabilities in clinical practice.

The developed model achieved promising results in the diagnosis task, effectively addressing the limitation posed by the small dataset. However, it is worth noting that the performance could potentially benefit from further improvements with the utilization of a larger-scale dataset for both training and validation tasks. Additionally, the current model focuses on the diagnosis of one fracture type per X-ray image, and its capability to provide multiple types simultaneously may require additional refinement.

5.7 Summary

This chapter presented a novel CNN approach for PFF diagnosis, tackling challenges related to dataset limitations, including size constraints. It defined the clinical process for diagnosing PFFs X-ray images, emphasizing the regions that clinical specialists focus on. The approach incorporates clinical expertise to guide feature extraction and fusion from the most distinctive regions within THR X-ray images, all within a multi-task CNN architecture.

Several experiments have been carried out to show the effectiveness of the developed approach. These experiments demonstrated that combining clinical knowledge has enhanced the overall

outcomes of fracture classification. Results show that the developed method is accurate with an overall 89.8% AUC. This outperformed the state-of-the-art results achieved using ResNet, which had an AUC score of 81.5%.

Chapter 6

Reassembly of Fractured Object Using Fragment Topology

6.1 Introduction

With the current development of 3D modelling techniques, the acquisition and processing of 3D objects have significantly improved. This also has influenced the reassembly of broken objects that has direct applications in various scientific domains, such as computer-assisted surgery, digital heritage archiving, forensic evidence analysis and archaeological reconstruction. For instance, CAP for bone fracture reduction procedures is used to assist orthopaedic surgeons in planning for fracture surgery to identify the correct matching of the bone fragments, especially in complex fracture cases. On various domains of reassembling 3D objects, the problem is defined as the automatic process of identifying the fractured part of an object, searching for corresponding fragments from the fractured object set and finally matching these parts in a virtual representation to form the original object. However, matching between broken fragments can not be compared to matching in general 3D object processing. When the object breaks, new surfaces are generated which are commonly called fractured regions and the surfaces that were not affected by the fracture are called intact regions. Figure 6.1 presents these regions. The challenges in matching broken object lie in the patterns between correlated fragments that are matchable surrounding the fractured region only, while in general or partial 3D matching the overlap patterns are more notable. In addition, the erosion and missing some parts of the fragment will introduce significant complexity to the problem.

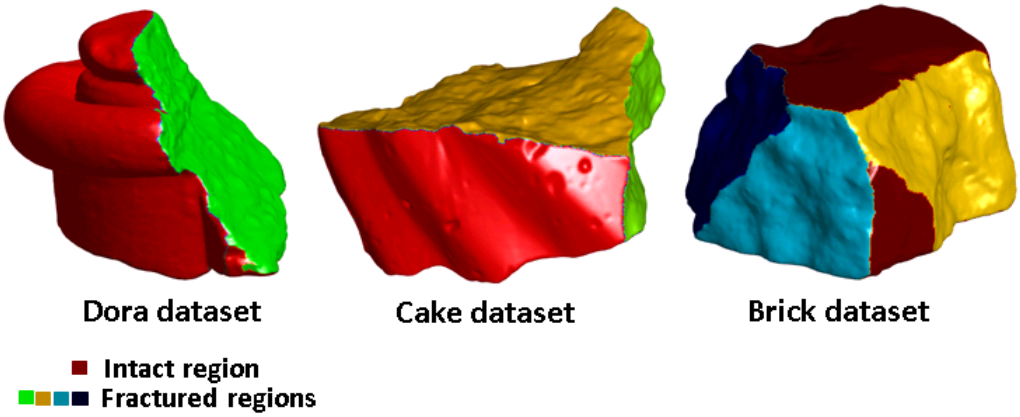


Figure 6.1: Examples of intact and fractured facets for different fragments.

The existing approaches for reassembling fractured objects can be divided into three main categories; reassembling based on fragment regions matching, reassembling using a template as guidance to reconstruct the object and methods that depend on expert support.

The region-based matching approaches commonly utilised the local surface properties to find similarities between adjacent fragments. Some approaches relied on the intact regions only to match fragments. For instance, Sagiroglu and Ercil 2006 used inpainting and texture synthesis methods to predict the pixel values in the band surrounding the fragment border. The texture features and confidence values were derived to compute an affinity measure of correlated pieces. The alignment of parts and assembly was determined by maximizing the affinity measure. FFT shift theory was utilized to find the alignment between pieces for the puzzle reconstruction. Indeed, intact-based matching approaches commonly assume that only the intact surfaces are acquired when scanning the puzzle pieces. However, such scan and clear segmentation between the intact and fracture regions might be difficult, especially when fragments are small. Other approaches considered the fracture regions to reconstruct the object. For thin shell objects such as a pot or fresco fragments, the fracture region is often treated as a boundary curve of the fragment and the problem is reduced to 2D puzzle solving (Kong and Kimia 2001; Gama Leitão and Stolfi 2002; Amigoni et al. 2003; J. C. McBride and Kimia 2003; Dellepiane et al. 2011). Kong and Kimia 2001 sampled the contour based on a polygonal non-uniform approximation in different scales and stored a curvature value for each point. Matching was accomplished based on a coarse-scale representation of curves, then refined using the elastic curve approach. The same approach was utilised by Gama Leitão and Stolfi 2002 but used a uniform sampling of the contour of the fragment. Then, a dynamic programming sequence matching algorithm was

applied to compare the curvature-encoded fragment contour at different scales of resolution. The correspondence between two outlines was defined as a collection of pairs of samples that allow many-to-one mapping between curve points. To discard the incorrect pairs, coarse fragment fracture curve representations were used. The results were refined using a geometrically smaller sampling stem until a final set of pairs was identified. Another method for 2D curve matching was introduced to address the problem of unconstrained curve matching which was considered to be computationally expensive. Many approaches constrained the matching by detecting critical features, such as corners, and using them to guide the search. For instance, J. C. McBride and Kimia 2003 examined the matches that begin at fragment corners only and then used curve-matching with normalised energy to determine how far the match extends. The multi-scale was also used to reduce the cost of matching operations. The corner points were extracted by locating the extreme curvature for all contours. Using these points, the matching of the sub-segments was performed on four scales. The coarse representation of contour i.e. the corner points and the matching was performed using elastic curve matching as described in (Sebastian et al. 2003). The second and third representations were acquired by re-sampling the sub-segment at even intervals of arc-length across the original contours. On the second level of representation, a group of the endpoints was determined, knowing the start point from a previous level, by balancing similarity and extent of the matching to get locally optimal sub-contours. The third level used the open curve matching method to obtain a more accurate measure of similarity. The fourth level represented the original set of points from the fragment's contour. On this level, the least squares method was performed to obtain the optimal registration. Finally, the pairwise affinity was measured by applying the following cost metric:

$$C_{total} = \lambda_1 C_{distance} + \lambda_2 \sqrt{C_{length}} + \lambda_3 \sqrt{C_{diagnostic}} \tag{6.1}$$

where $C_{distance}$ is the average distance between corresponding points, C_{length} measures the confidence that the match is correct based on the arc length and the $C_{diagnostic}$ measures the complexity of the common boundary.

For general 3D fragments reassembly, the pioneered work was introduced by Papaioannou and Evaggelia-Aggeliki Karabassi 2003 where the fragments were initially segmented into intact and

fractured patches and the pairwise matching was performed using the curvature difference metric. In another work, the authors combined both the surface and curve matching strategies (Papaioannou and Evaggelia-Aggeliki Karabassi 2003). The fractured surfaces were extracted and classified into external and internal surfaces based on whether one or more intact surfaces were adjacent to them or not. This determined which matching type can be applied to each fractured facet. For the internal case, unconstrained surface matching was applied (Papaioannou, Karabassi, et al. 2002), where two fragments patches were roughly aligned based on corresponding region normals and their relative pose was optimised according to the curvature difference. For the external fractured facet case, a boundary-constrained method was applied. Two boundary lines were compared using a signature based on discrete curvature and torsion as described in (Üçoluk and Toroslu 1999). The Euclidean distance of the local boundary feature was utilised to measure the matching. For the final alignment, a closed-form solution that operates on consecutive triads of corresponding boundary points was used. A set of N-2 transformations was obtained and the one that minimised the average distance between the two entire point sets was stored. Another approach for solving the free-form 3D puzzle problem computed multi-scale integral invariants for the surfaces to segment the fragment surface and define the matching between fragments (Q.-X. Huang et al. 2006). First, the fragments were segmented into a set of facets using the defined surface sharpness descriptor and classified into original and fractured facets using the computed surface roughness descriptor. Then, pairwise matching between fragments was defined using features clustered for each fractured facet. Coarse alignments were performed for all pairwise matching using the Forward Search Method to define the initial alignment. The results were further improved using ICP algorithm. Although this method showed proper assembly for the presented cases, it is notably complicated and involves numerous specialized algorithms for segmentation, classification, multi-scale feature extraction, pairwise matching, registration and collision detection. This complexity poses challenges for implementing and adopting the system. In addition, small fragments might not produce an accurate matching due to difficulties in the segmentation of these fragments.

Rather than relying on feature computation to define the matching score between surfaces, Mavridis et al. 2015 introduced a three-level coarse-to-fine search strategy that was based on the residual distance between fragments.

Approaches of the second category of the fragment reassembly reconstructed the fragments by

finding their best match to a template model using the intact surface information. Fürnstahl et al. 2012 used a reference of a healthy bone template for humerus fracture to aid fracture reduction surgery. The initial matching of the fragments to a contralateral bone model was performed by identifying candidate points, which represented the outer surface of the bone. These points were defined based on the local normal vectors. The final alignment between fracture surfaces was obtained using the ICP method. However, the fracture lines were obtained manually in this approach. Yin et al. 2011 proposed a two-step framework for assembling skull fragments. The first step matched each fragment to a template skull and roughly reassembled the object using the ICP-like template integrated with the slippage features and spin-image descriptor. The second stage refined the assembled fragments by analysing the break curves. Similarly, Yu et al. 2012 developed a method that matched skull fragments to a template but by using heat kernels and the RANSAC algorithm for matching, refinement and assembly calculation. K. Zhang et al. 2015b developed a reassembling method that combined both matching of adjacent fragments' surfaces and guidance from a template. A model with similar geometry to the original object was used as a template. A partial matching between fragments and a template is determined by comparing feature points. The Signature of Histograms of Orientations was utilised as a point descriptor (Tombari et al. 2010). The matching was refined using the Random sample consensus (RANSAC) method (Fischler and Bolles 1981). For pairwise matching of fractured regions, a curve matching strategy was used, which might not be sufficient for thick fractured objects. These methods mainly depended on the availability of the template model and could not construct a general model.

In the interactive expert-based approaches, user involvement for manual guidance is essential for reassembly methods. P.-Y. Lee et al. 2014 introduced a method that required a manual selection of several pair points on the fracture surface outlines of the two adjacent fragments to compute the alignment. A coordinate transformation algorithm (Lai and K.-J. Chen 2007) was used for the alignment of the fragments. Beibei Zhou et al. 2009 and Willis et al. 2007 developed an interactive system that enabled the user to specify the initial matching between the fragments. The system initially segments each bone fragment into intact and fractured surfaces based on the analysis of the bone density. After the selection of correspondences between fractured regions, the modified ICP algorithm was applied to perform the final alignment. Mellado et al. 2010 proposed a real-time interaction loop system that enabled the user to approximate the initial position and orientation between two fragments and continuously correct and validate the pose

using the Kd-tree and ICP algorithm.

Many objects break into a large number of pieces which necessitates extensive efforts for their reassembly. As a result, user intervention in this complex process could prove to be both time-consuming and susceptible to errors. On the other hand, previous work for automating the reassembly process attempted either to match the fragments to a template model or to define relations between fragments using fragment geometries. These approaches focus on different regions of the fragment model i.e. intact and fractured regions. The template-based approaches used the intact regions of the model to locate the fragments in a template model. However, the optimal template model of the broken object is not always available. On the other hand, the process of assembling a fractured object using its fractured regions relied on the existence of salient features on the broken side. However, defining these salient properties can be challenging, and might not be generated when the object is broken. Consequently, this leads to a large number of potential matching fragments and substantially increases the complexity of finding the correct correspondence pairs.

This chapter aims to tackle the above problems by introducing a new representation of fragments. This representation is inspired by the manual assembly of broken objects. When an expert attempts to reconstruct the original broken object, his main focus is on the fractured boundary of each fragmented piece. Additionally, the surrounding intact area of the fractured boundary might provide further clues about the corresponding fragment. From these observations, a fragment representation is defined by decomposing the fragment's surface into intact and fractured facets. Each facet is represented with a graph node and two nodes are linked with an edge if they are spatially adjacent. This representation is referred to as a fragment topology , see Figure 6.2 (b). This topology simplifies searching for potential matching fragments using both fractured facets and their adjacent intact properties. In addition, the orientation of the matching fragment is easy to identify using the adjacent intact facets information of each node i.e. matching fractured boundary and intact boundary. To extract fragment topology, the fragment is segmented into intact and fractured facet using a new defined surface feature which improves the facet segmentation. Then, an iterative matching and representation of fragments are used to find the optimal matching pairs and form the original object. The contributions of this work are: (1) this chapter proposes an integrated approach that combines the intact facet properties with the fractured boundary curves to assemble fractured object and minimize

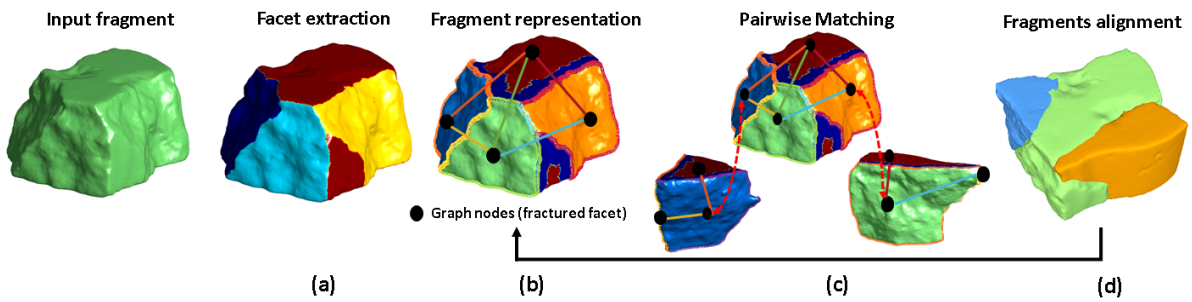


Figure 6.2: The general workflow of the proposed method. (a) segmentation of fragment surface into intact and fractured facets.(b) Creation of fragment topology and descriptors. (c) Matching between graph nodes to find possible matches (d) Iterative Matching and representation of fragments.

the potential matching. (2) a fragment topology is introduced which represents the fragment features and simplifies the search for possible matching, further enhancing the reassembly process. (3) a new feature is proposed in this work that improves the facet segmentation and classification process significantly.

6.2 Assembly Approach

This work introduces a novel, fully automatic method for reassembling fractured objects. The fragment topology is proposed for representing the fragment's surface based on its part arrangement. This topology is used to guide the search for possible matching fragments.

Given a set of fragments F_i, optimal matches are identified to reconstruct the original shape of the object. The reconstruction of the broken object has four main steps as illustrated in Figure 6.2.

1. Segmentation of fragment surfaces into intact and fracture facets (Section 6.2.1).
2. A graph representation of the fragments (Section 6.2.2).
3. Extraction of boundary curve features and other properties to measure the pairwise matching between fragments (Section 6.2.3).
4. Iterative assembly of the object by matching the selected corresponding pairs and updating the representation with the combined fragments (Section 6.2.4).

6.2.1 Fragment segmentation

The initial step in matching fragments is the analysis of the fragments' surfaces and extraction of regions of interest. When the object breaks, new surfaces are generated that form a fractured region. Therefore a fragment's surface can be categorised as intact and fractured. Each of these regions provides different characteristics that can support finding correct matching between the fragments. The intact regions identify the continuity of patterns and geometries between fragment surfaces, especially on the boundary areas close to fractured regions. On the other hand, the fractured regions define complementarity mating between pieces.

The segmentation process was performed to divide each fragment surface into distinct regions to avoid the wrong matches and reduce the computational effort. The primary goal in this step is to extract segments for designing a topology. Each segment should include either the intact or fractured surface but not both.

The proposed workflow started with an initial segmentation phase using the Region Growing approach (Mehnert and Jackway 1997) with specific criteria. This step is followed by a merging phase that includes the classification of regions into fractured and intact.

Region Growing

Given a 3D mesh of a fragment, it starts by selecting a random face from the surface as a seed element and grows into a region by iteratively adding neighbours based on specific conditions. Let S be the seed element of the current segment and L be all unassigned neighbouring faces. The compatibility score D_{l_i} is calculated based on the angle between the normalized average normal of the seed element and its 1-ring neighbouring face (n_s) and the normal vector of the examined face (n_{l_i}):

$$D_{l_i} = \cos^{-1}(n_s.n_{l_i}) \ . \tag{6.2}$$

The local average normal n_s is:

$$n_s = \frac{\sum_{k \in S_r} n_k}{\left|S_r\right|} \ . \tag{6.3}$$

where n_k is the k^{th} face normal within 1-ring neighbouring of seed element S_r.

The above method worked well on a planer surface but resulted in multiple segmentation on a fractured surface due to the presence of highly irregular surfaces (see Figure 6.3 (a)). Therefore, a post-processing step is required to improve the segmentation accuracy.

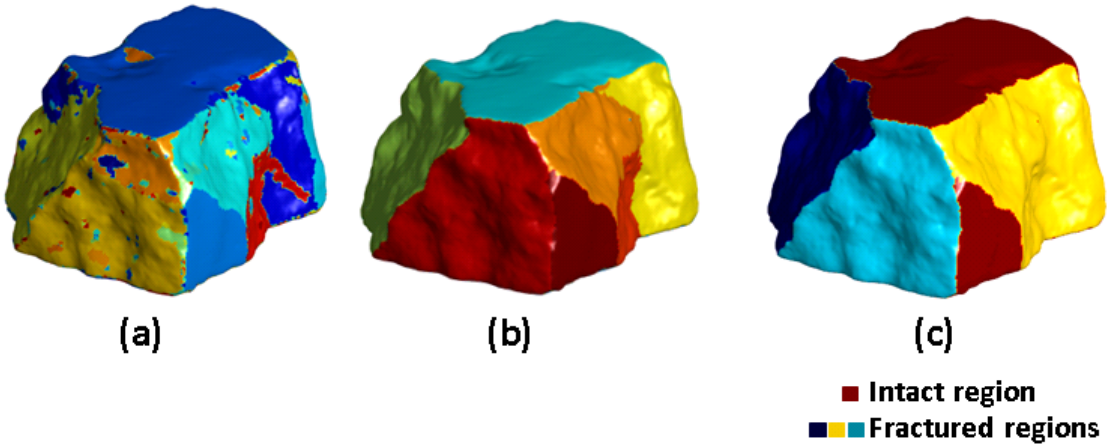


Figure 6.3: Segmentation step (a) Region growing (over-segmentation problem) (b) Merging criterion (segment area) (d) Merging criterion (segment type + compatibility score)

Region Merging

The region-growing step provided an initial rough segmentation of a surface. In this phase, the segments were categorized based on two criteria: the surface area of the segment and the type of surface (fractured or intact). Rather than merging the segments and then classifying them as in the existing approaches (Q.-X. Huang et al. 2006; Mavridis et al. 2015; Thomas et al. 2011), the type of segment was exploited to perform the merging.

The first merging criterion iteratively combined all small segments i.e. the segment where the area is less than a threshold, to adjacent large segments based on a minimum compatibility score. The compatibility score was determined as follows:

$$S_{l_i} = \min\left(\cos^{-1}(n_{s_l}.n_{s_s})\right) \ . \tag{6.4}$$

where n_{s_l} is the average normal vector for the large adjacent region and n_{s_s} is the average normal vector for the small region. Figure 6.3 (b) demonstrated an example of applying the first merging step.

The second merging criterion considered merging segments of the same type and within the compatibility score, Equation (6.4). Broken objects vary in terms of the intact surface. Fragments with planar intact surfaces are easy to classify, as opposed to, fragments with highly curved intact surfaces or detailed patterns. Therefore, a new feature was defined to distinguish between intact and fractured facets. The ratio of surface curvature for each segment was defined and compared to the average curvature of the fragment surface. For each fragment, the average

of maximum curvature (ϵ) is calculated as:

$$\epsilon = \frac{\sum_{k \in P} C_{max_k}}{\left|p\right|} \quad . \tag{6.5}$$

Where C_{max_k} is the maximum curvature for point k in the fragment and $\left|p\right|$ is the number of points on the fragment. ϵ is used to define the ratio of maximum curvature in each segment as:

$$R_s = \frac{\left|C_{s_i} \geq \epsilon\right|}{\left|P_s\right|} \quad . \tag{6.6}$$

where C_{s_i} is the maximum curvature of points P_s on segment s_i. The resulting segment ratios (R_s) were clustered into two groups using the k-mean algorithm. Clusters with a larger centroid value represent fractured segments.

6.2.2 Fragment Representation

This work introduced a fragment topology to simplify fragment surface representation and guide the search for optimal pair matching. The matching fragment will have a similar topology in terms of intact and fractured facets. The seek is to filter the wrong potential matches by defining simple properties along the fragment topology and find out how this can reduce the search for potentially matching fragments.

For each fragment, given the segmented facets from the previous step, a graph $G = (V, E)$ was defined where each node $n_i \in V$ is denoting a segmented facet f_i associated with a type of facet (intact or fractured) and the boundary curve of the fractured facet. The extraction and description of the boundary curve will be illustrated in Section 6.2.3. The edge $e_{i,j} \in E$ between nodes n_i and n_j denotes the boundary connecting two facets associated with attributes such as the arc length of the boundary curve connecting two facets, and the start and end points of the boundary curve. Figure 6.2 (b) shows an example of the fragment representation.

6.2.3 Feature Extraction

Boundary Curve of Fractured Facet

As illustrated before, when trying to reassemble the broken object, the focus is on the boundary of the broken facet. Accordingly, a measure of pairwise similarity was defined based on these boundaries. Multi-views of a fragment were used to extract 2D boundary curves. The fractured

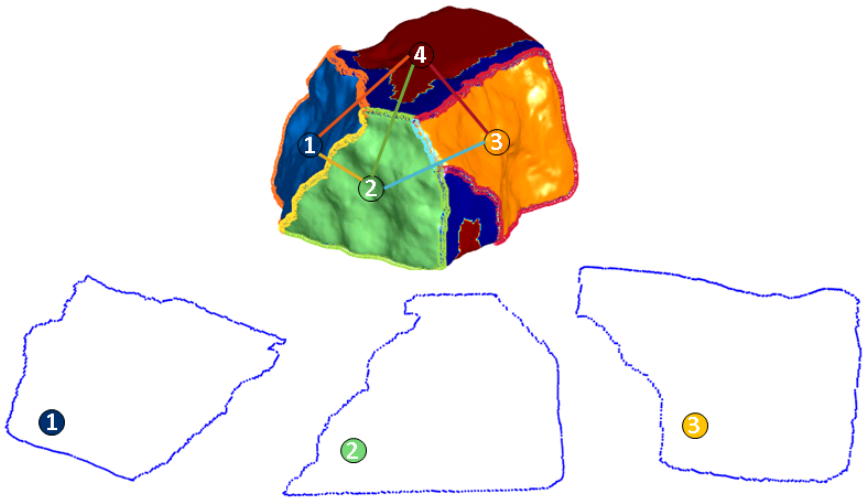


Figure 6.4: Examples of extracted 2D boundary curves.

facet was identified in the previous steps. The centroid point of a fracture facet used as a viewpoint to project the facet boundary into the XY plane. The number of viewpoints is based on the number of fractured facets. Figure 6.4 shows an example of the extracted boundary curves.

Boundary Curve Descriptor

The extracted boundary curve is highly affected by noise and modelling variations. Also, the matching boundaries might be extracted in different orientations. To mitigate these problems, boundary curves were described with a robust descriptor that is invariant to scale and rotation. Fourier descriptors showed efficient descriptions for shapes that include scale, rotational and translation invariant (Gonzalez Rafael 2008). The boundary points are first converted into complex numbers to extract the curve signature, as in the following:

$$S(k) = X(k) + jY(k) \quad . \tag{6.7}$$

where $X(k)$ and $Y(k)$ represent x and y coordinates of the curve. Fast-Fourier transform of the boundary signature provides a Fourier descriptor and is defined as:

$$a(u) = \sum_{k=0}^{K-1} s(k) e^{-j2\pi uk/K} \quad . \tag{6.8}$$

where $u = 0, 1, ..k - 1$.

In order to achieve translational invariance, the DC component of the Fourier descriptors was set to 0. For scale invariance, all the coefficients of the Fourier descriptor are normalised by the second coefficient:

$$a(0) = 0, \;\; a(u) = \frac{a(u)}{a(1)} \;\; . \tag{6.9}$$

Both magnitude and phase values were considered, however, phase values were affected by rotations and start point variations. For this reason, the topology representation of fractured facets was used to define rotations and starting points of the boundary curve.

6.2.4 Matching

For reconstructing the final object the following steps were followed:

- defining the whole-to-whole matching between the fragments.
- combining the matching fragments group as one fragment.
- recomputing the new representation of the combined fragments.
- iteration of the matching and representation until the final assembly. of the fragments is found.

Pairwise Matching

The process of finding potential match fragments list was simplified by integrating simple properties based on the defined representation. The typical way of searching for the potential matching fragment is based on finding similarities between extracted feature points, which results in a large set of wrong potential matching points. This is typically followed with a refinement step such as using the RANSAC algorithm (Son et al. 2018). On the other hand, this work introduced searching for potential matching fragments using several factors: The topology of the fragment, for example, a number of intact and fractured neighbours and the arc length of the boundary curve connecting intact and fractured facets, the area of the fractured facet and boundary of fractured facets. The above geometrical properties are used to reduce the potential matching pairs.

In order to identify exact matching pairs, a similarity score between possible pairs was defined as the Euclidean distance of their Fourier descriptors of the boundary curve.

Dataset	Model name	Type	# fragments	Vertex range
Vienna	Brick	stone	6	70k - 111k
	Cake	mortar	11	57k -151k
	Sculpture	clay	7	95k - 198k
PRESIOUS	Nidaros Crypt Tombstone	stone	5	110-150
	Nidaros Cathedral Column Base	stone	5	40k - 70k
Bone Model	Femur	Foam cortical shell	2	57k - 64k

Table 6.1: Fractured objects used to evaluate the facet extraction methodology

Group Matching

Based on the fragment topology similarity and the defined similarity score, the best matches between fragments were selected and the assembly graph was initiated. All fragments were encoded with their extracted best matches in a reassembly graph $G = (V, E)$. Each node $n_i \in V$, denotes a fragment F_i and each edge between the nodes n_i and n_j denotes the correspondences between the fragments. Each connected component in the graph was considered as a whole fragment. So its new representation and pairwise matching were recalculated and the reassembly graph was updated until all graph nodes were connected.

6.3 Results and Discussions

To validate the proposed method and the effectiveness of each step, first, the facet extraction step is validated and then the fragment assembly. The evaluation settings and the results of each experiment are demonstrated in the following sections.

6.3.1 Facet Extraction

The proposed facet extraction method was assessed on four kinds of fractured objects; stone, mortar, clay and foam cortical shell. Table 6.1 showed an overview of the 3D models used to evaluate the proposed facet extraction method. The objects used in this experiment were broken into different numbers of fragments and various sizes and shapes. Also, some fragments were exposed to weathering and erosion which changed surface properties. The bone models were generated from CT scans.

The input to this method is a digital model of broken fragments. The tested number of vertices for each fragment was between 40k to 151k and the total number of tested fragments was 36 fragments. Figure 6.5 illustrated some examples of the segmentation phases for each of the used three datasets.

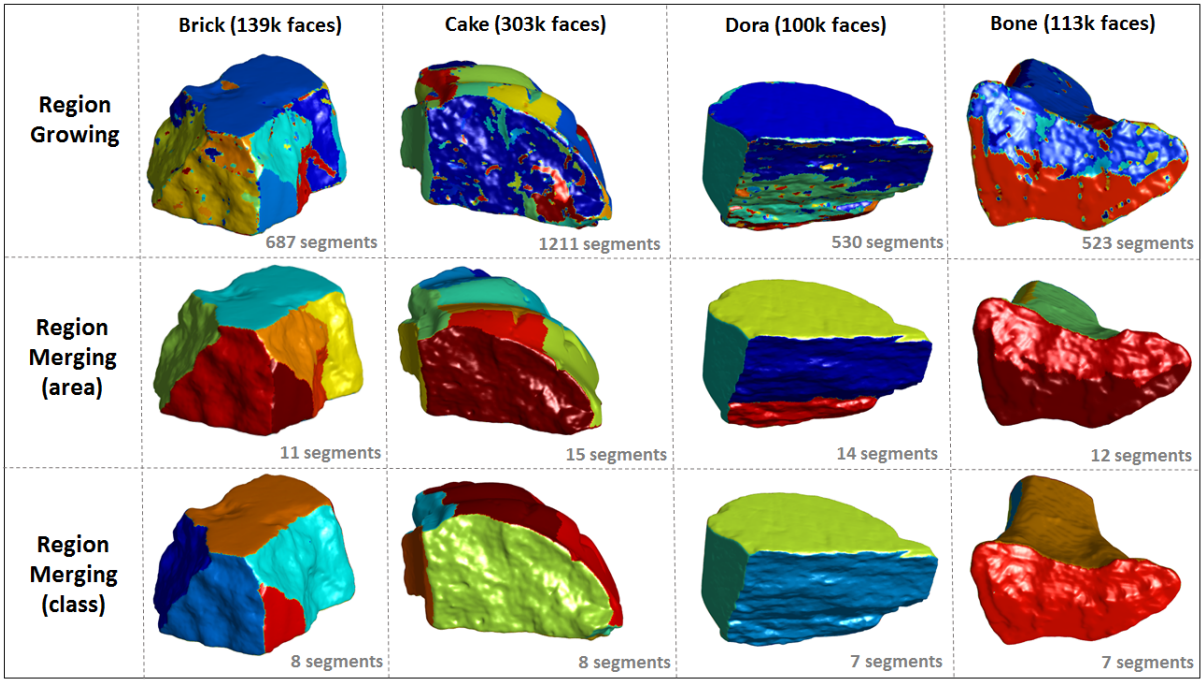


Figure 6.5: Result of facet extraction in each step for different datasets (see Table 6.1.)

The evaluation of the facet extraction step examined the performance of the segmentation of facets and the classification of each facet as either intact or fractured. For the segmentation, the manual segmentation of the fragment surface was used as ground truth. The Dice Coefficient (DC) was used as an evaluation metric for the segmentation. It computed the spatial overlap between two sets of segmentation i.e. the ground truth and the resulting segmentation. The DC score ranged between 0 (not similar) and 1 (similar) and is defined as follows:

$$DC(R_a^i, R_g^{i_t}) = \frac{2 \times \left|\left|R_a^i \cap R_g^{i_t}\right|\right|}{\left|\left|R_a^i\right|\right| + \left|\left|R_g^{i_t}\right|\right|} \quad . \tag{6.10}$$

Where R_a^i and $R_g^{i_t}$ are the automatic and manual segmentations, respectively, and i_t is the index of the closest segment from the ground truth segment (S_g) to R_a^i which defined as:

$$i_t = argmax_k(\left|\left|R_a^i \cap R_g^i\right|\right|) \quad . \tag{6.11}$$

The DC between two segmentation the result segmentation (S_a) and the ground truth segmentation (S_g) is computed as follow:

$$DC(S_a, S_g) = \frac{\sum_{i=1}^{k} DC(R_a^i, R_g^{i_t})}{k} \quad . \tag{6.12}$$

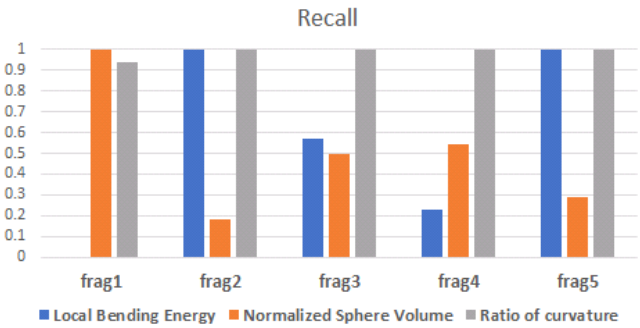


Figure 6.6: The Recall value for Local Bending Energy (Q.-X. Huang et al. 2006), Normalized Sphere Volume (Mavridis et al. 2015) and the Ratio of curvature.

To evaluate the efficiency of the facet classification model, the overall accuracy was measured as the ratio between correctly predicted facet type and the total number of facets. The results were further assessed using Recall to measure the rate of fractured facets that were classified correctly and is defined as:

$$Recall = \frac{\left|\left|S_a^f \cap S_g^f\right|\right|}{\left|\left|S_g^f\right|\right|} \quad . \tag{6.13}$$

Where S_a^f represents the facets that are classified as fractured and S_g^f is the ground truth of fractured facets.

Results - Figure 6.5 showed the results of the facet extraction steps for different types of fragments. The first row presented the Region Growing algorithm results with ($D_l = 5°$). It performed well on the flattened surface, whereas provided over-segmentation in the curved regions, which resulted in a 0.01 DC score. Using merging by area only has enhanced the segmentation results to a 0.66 DC score. However, when both merging criteria were applied i.e. the area and the class, the segmentation was improved and provided a 0.87 DC score (for area threshold $< 2\%$ and $D_l = 5°$). Choosing a good threshold was a difficult problem, therefore it was selected empirically.

The introduced feature for classifying fractured facets achieved an 82% overall classification accuracy. Notably, 97% of the fractured facets were correctly classified, demonstrating the effectiveness of the proposed features in accurately identifying these critical regions. Figure 6.6 presented the Recall value for extracting fractured facets utilizing the ratio of curvature compared to existing state-of-the-art method techniques. Q.-X. Huang et al. 2006 computed the local bending energy for each point to classify the segmented region and Mavridis et al. 2015 utilized the normalized sphere volume as a classification factor. The Nidaros Cathedral

dataset was used to compare the results. Even though Q.-X. Huang et al. 2006 illustrated enhanced performance in frag2 and frag5, using ratio of curvature improved the fracture facet classification significantly with a recall value of 100% in frag2, frag3, frag4 and frag5. It is also important to note that both Q.-X. Huang et al. 2006 and Mavridis et al. 2015 methods required expert intervention to adjust multiple parameters for segmenting and classifying the fractured facet, the proposed approach achieved better outcomes without user intervention. Additionally, the proposed method was assessed on various types of object materials and datasets, while Mavridis et al. 2015 method appeared to be effective for specific fractured objects.

6.3.2 Fragments Assembly

Evaluating the reassembly of the broken object method is restricted by the lack of the original model. Therefore, the proposed method was validated using both simulated and real models. For the simulated models, different 3D models were created and shattered into different configurations. The creation of the simulated 3D models was made using Autodesk Maya software. Refer to Figure 6.7 (a) as an example. First, a shatter effect was applied to a 3D model. Then, random translations and rotations were applied to each fragment to simulate the occurrence of a fracture. In addition, the approach was evaluated using real broken objects provided in Q.-X. Huang et al. 2006. Figure 6.7 (b) presented a brick fractured example.

To assess the reassembly result in comparison to the original model, it was necessary to exclude the fractured regions. This task required accurate segmentation, which presented challenges and was prone to errors. Instead, the evaluation was performed by comparing the assembly of the original fragments with the resulting assembly. For the simulated dataset, the ground truth was generated by breaking the object and preserving the connected fragments. Subsequently, random transformations were applied to simulate the shatter effects. Figure 6.7 (a) showed the different stages for generating the simulated dataset. For the real fractured object, manual aligned fragments were utilized as the ground truth for evaluation.

The accuracy of the reassembly is computed using the error measure (E_{ref}) which defined as the root mean square error between the reference and reassembled pieces.

$$E_{ref} = \sqrt{\frac{\sum_{i=1}^{N}(F_{r_i} - F_{a_i})^2}{N}} \quad . \tag{6.14}$$

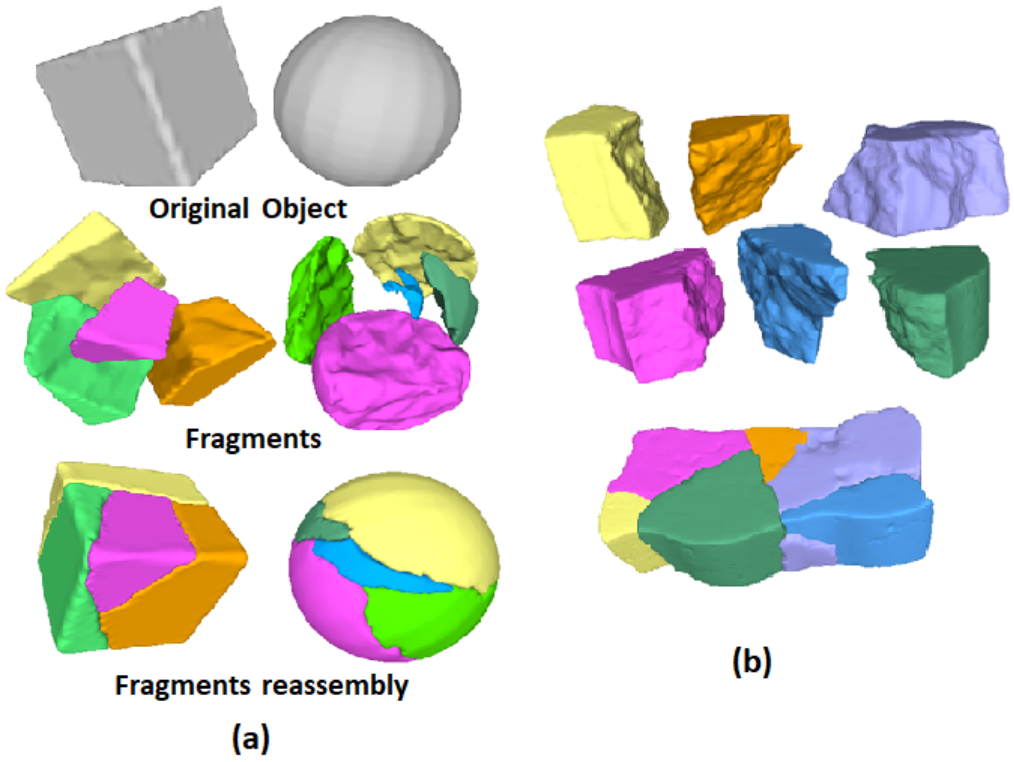


Figure 6.7: (a) Example of the simulated dataset. (b) Brick fractured model (Q.-X. Huang et al. 2006).

Where F_{r_i}, F_{a_i} are the i^{th} point on the reference and aligned model respectively and N is the number of the points.

Results - The proposed method was ran on desktop with 3.60 GHz Core i7 CPU and 16 GB RAM. Table 6.2 showed the run-time of the reassembly method. The total computation required 14 seconds for 6 fragments object and might increase depending on the number of fragments. In the reassembly process, the most time-consuming step was the search for potential matching fragments. For brick fragments, the K. Zhang et al. 2015a method required 2 seconds to set the potential matches and Son et al. 2018 required about 16 seconds. In contrast to these methods, the proposed method found matching fragments of a brick model within 0.7 seconds. This was decreased by introducing the topology of the fragment, which restricted the number of possible matching combinations.

Figure 6.8 and 6.9 showed the reconstruction of fragmented object 1 and object 2. These objects were shattered into different configurations and composed of fragments that have a partial relationship to each other. The proposed method can identify the initial matching pieces correctly and reconstruct the final shape effectively. The pairwise alignment between the

Model	#V	#F	t_{rep} (s)	t_{pm} (s)	t_{mm} (s)
obj1	80k	3	1.34	0.1055	0.01
obj2	108k	4	2.3	0.32	0.07
brick	534k	6	13	0.7	0.13

Table 6.2: Performance of the proposed method: (model name, number of vertices of all fragments, number of fragments and time in seconds for representation process of all fragments (t_{rep}), create potential matching (t_{pm}) and multi-piece matching(t_{mm}))

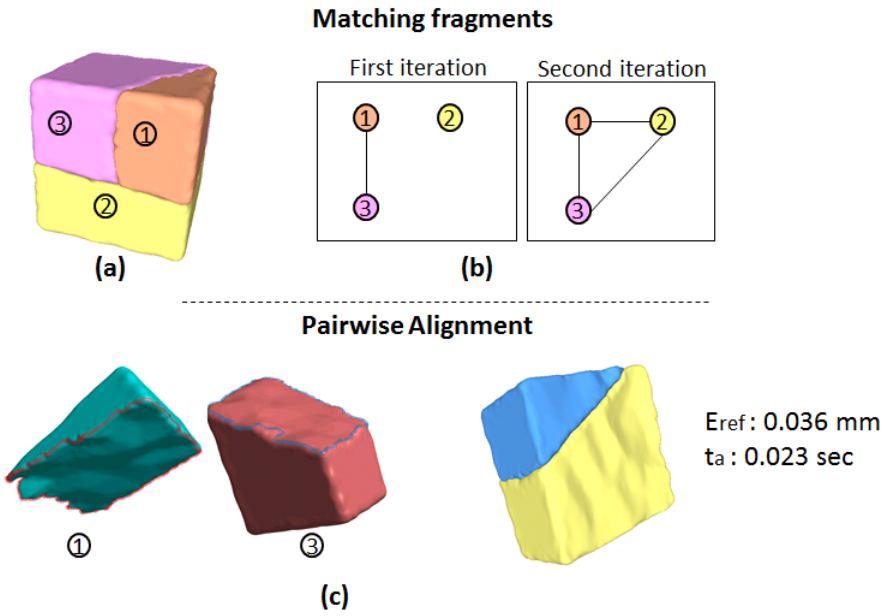


Figure 6.8: Object1 model: (a) Multi-piece reconstruction of the fragments. (b) iterations to reconstruct the original object. (c) Pairwise alignment, the resulting reference error (E_{ref}) and the alignment time (t_a).

fragments was measured based on the matched fractured boundary using the ICP method and provided efficient and accurate alignment with average error $E_{ref} = 0.047$ mm. Figure 6.8 (c) and 6.9 (c) showed examples of the pairwise alignment between two fragments.

In addition, the proposed method was tested on a real model (brick model) which was fractured into six fragments. The broken brick model was affected by erosion and each fragment can be matched with one or more of the other fragments. Figure 6.10 (a) illustrated the multi-piece matching construction between the brick fragments. The introduced topology representation can achieve 0.19 mm matching accuracy after two iterations, see Figure 6.10 (b).

The previous reassembly methods, Q.-X. Huang et al. 2006 and Son et al. 2018, proposed a complex descriptor of fractured facet that required a large number of discriminating points in order to accurately describe the facet. In both these methods, the potential pairwise matching process resulted in incorrect matches that required further refinement steps to reduce the possibility of

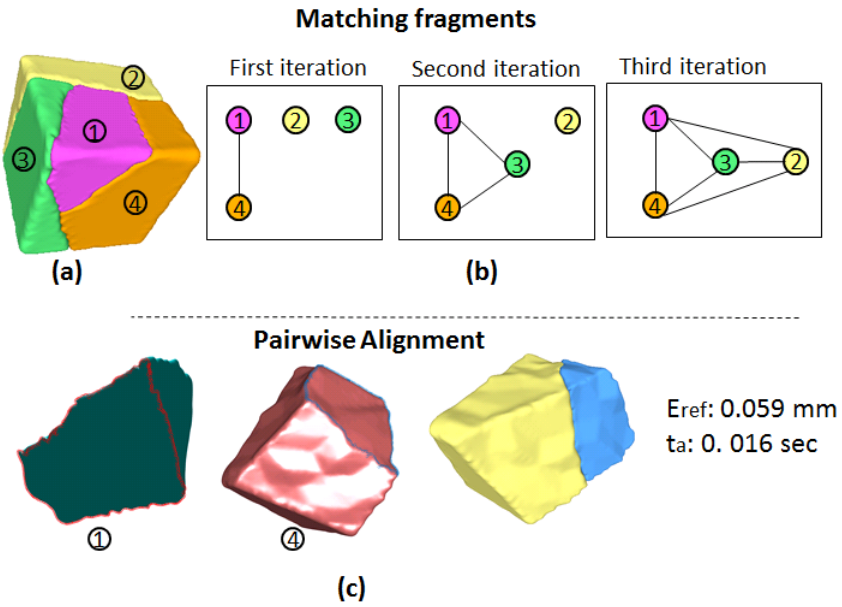


Figure 6.9: Object2 model: (a) Multi-piece reconstruction of the fragments. (b) iterations to reconstruct the original object. (c) Pairwise alignment, the resulting reference error (E_{ref}) and the alignment time (t_a).

wrong matches, leading to the increased complexity of the algorithm and the matching time. The proposed fragment topology combined the fractured region boundary and its relation to adjacent intact facets to define matching fragments. This simplified the searching for matching fragment and provide promising results.

6.4 Summary

This chapter developed a new approach for solving 3D puzzle problems directly applicable in many fields including computer-assisted surgery and archaeological reconstruction. The developed approach combined intact facet properties with the fractured boundary curves to assemble the fracture and minimize the potential matching. A novel representation called 'fragment topology' is introduced which captures features of fractured facets and their relationships to intact facets. In addition, a new feature 'ratio of curvature' is proposed for segmenting and classifying fragment facets.

Several experiments have been conducted to validate the effectiveness of the developed approach. These experiments demonstrated that using fragment topology has simplified the search for potential matches and enhanced the reassembly process. Unlike the existing approaches such as Q.-X. Huang et al. 2006 and Son et al. 2018, this approach avoids the complexity and increased

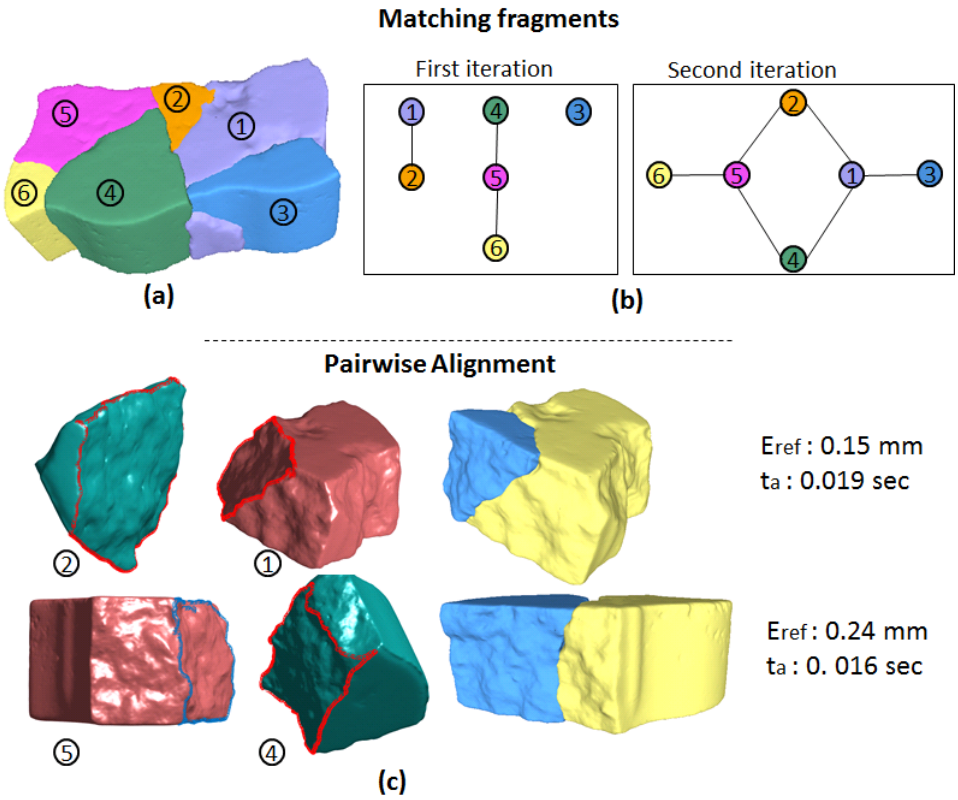


Figure 6.10: Brick model: (a) Multi-piece reconstruction of the fragments. (b) iterations to reconstruct the original object. (c) Pairwise alignment, the resulting reference error (E_{ref}) and the alignment time(t_a).

computational cost associated with the use of numerous specialized algorithms in each step. Instead, this is achieved through the proposed fragment topology that represents a fragment as graph and constrains matching fragments using simple features like adjacent fracture types and fractured facet areas.

The experiments also showed that using the ratio of curvature has improved the facet classification and segmentation results significantly. The proposed feature provided an accurate classification of fractured facets by identifying 97% correctly in the test dataset. Using Nidaros Cathedral dataset for comparison, this work achieved 98.6% recall score. This outperformed the state-of-the-art results achieved using Q.-X. Huang et al. 2006 and Mavridis et al. 2015, which had a recall score of 56% and 50%, respectively.

Chapter 7

Conclusion

7.1 Summary of Main Outcomes

The focus of this thesis is on the design of the framework for automatic and reliable THR X-ray image diagnosis, enabling the analysis of intricate information and other features of interest and assisting in treatment planning. Many objectives were outlined in this context. These objectives were addressed as follows:

- Generation of THR dataset was achieved by collecting a large THR dataset from multiple trauma centres in the United Kingdom and a published dataset (Patel et al. 2021). The generated dataset involved a wide range of implant designs, post-THR X-ray images and different kinds of PFF fracture X-ray images. In addition the images were of different orientations and various positioning of the hip that could include the whole femur or partial part of the femur. The images were annotated by two clinicians for fracture labelling and one clinician for defining the fracture region boundary and Gruen landmarks positions. (Chapter 3 and Chapter 4).

- Automatic detection of implant landmarks of interest and the segmentation of the implant were addressed by integrating the prior shape knowledge with DL model. First, the shape of the implant component is represented by applying SSM based on the Gruen zone landmarks. Then, a hybrid approach for simultaneous regression of SSM parameters and segmentation maps of the implant component is developed. The final landmarks were computed by aligning the shape with the segmentation maps. It was found that

the incorporation of shape knowledge enhanced the results of both segmentation and landmarks detection. (Chapter 4).

- Developing a framework for diagnosing PFF images that addresses current data limitations and associated challenges was achieved by incorporating the clinical diagnosis process into DL. The clinical process of reading PFF X-ray images is defined first, and then these steps are utilized to design a knowledge-guided framework for PFF diagnosis. The framework consisted of four main feature extraction components and a joint fusion learning component that extracted the most distinctive regions of femur X-ray image enabling us to acquire comprehensive deep feature knowledge from the images and resulting in a more robust and effective solution. It was found that integrating clinical interpretation knowledge has improved the PFF diagnosis with a higher mean F1 score compared to the state-of-the-art model. (Chapter 5).
- Design a framework for the reassembly of the fracture object to assist in the pre-operation planning for fracture reduction was initially addressed by introducing a framework that combined intact and fractured facets boundary to reduce the possible matching between fragments. In addition, a new representation was introduced to represent a fragment and its associated features which were used to simplify the search for possible matching. The segmentation and classification of the fragment facets were improved by the proposed feature, a ratio of curvature. (Chapter 6).

7.2 Contribution to the Knowledge

The current practical methods for diagnosing and planning treatments for THR complications rely on interventions from clinical experts. However, this intervention can result in outcome variability, human errors, and potential diagnostic delays, especially in situations where there is a shortage of experienced clinical specialists and with a growing demand for THR treatments. This thesis developed approaches for automating and improving the diagnosis and planning of THR complications. This section emphasises the significance of the proposed approaches in advancing knowledge of the medical image analysis domain particularly in THR X-ray images. The scientific contributions of this thesis can be summarized in the following:

- Developed a DL-based method for automatically detecting, diagnosing, and localizing

PFF in THR X-ray images using CNN-based models like DenseNet and ResNet. To this end, a large database of THR images with PFF cases was generated associated with the annotations of fracture class labels and boundary boxes of the fracture region. In-depth assessments of numerous CNN-based architectures demonstrated that DenseNet achieved the best PFF detection performance, with F1 score of 95%. Challenges were identified in classifying fracture types, with typical CNN models achieving a lower F1 score of 54% as the task complexity increased. The fracture region localization methods were also assessed, with the CAM method providing an approximate visualization and Faster RCNN achieving a localization accuracy of 74.5%. Using the ROI as a pre-processing step for classification provided similar performance to using the entire image, a contrast to other fracture classification methods designed for specific anatomical regions like the proximal femur. This highlighted the challenges and pattern variations presented in PFF X-ray images. A journal article has been published to demonstrate this work (Alzaid, Wignall, et al. 2022).

- A novel CNN approach is introduced for segmenting the implant femoral component and predicting the Gruen landmarks, essential for THR radiograph analysis and detecting complications such as PFF and implant loosening. The Gruen landmarks are precisely defined to represent the implant's femoral shape, forming the basis for an SSM. A hybrid CNN network is developed to perform simultaneous implant component segmentation and SSM parameter regression, allowing for accurate implant landmark localization. An annotated THR image dataset is created for future research. Extensive experiments showed the approach's effectiveness with segmentation results of 80% dice score and HD of 8.8 px, outperforming the state-of-the-art result of 74% dice score and HD of 16 px. Likewise, landmark localization achieved a remarkable NMRSE of 0.33 pixels, surpassing results obtained with DenseNet 2.78 px. These contributions enhanced THR image analysis and provided valuable knowledge for clinical applications. A journal article has been submitted to demonstrate this work and currently is under review process (Alzaid, Lineham, et al. 2023).

- A new CNN model was developed for diagnosing PFFs, addressing dataset limitations and highlighting clinically significant regions. This involved observing and defining a clinical diagnosis process and automating the localization and cropping of Gruen zones

using detected landmarks. A multi-task CNN architecture was designed according to these clinical processes, enhancing the extraction of discriminative features in THR X-ray images to improve PFF diagnosis. Experimental results demonstrated the approach's effectiveness, achieving an overall AUC of 89.8%, surpassing the state-of-the-art AUC score of 84% that is achieved using ResNet. This work is accepted as a full paper in the 29th IEEE International Conference on Mechatronics and Machine Vision In Practice.

- A new approach to solving 3D puzzle problems is introduced, with broad applications including computer-assisted surgery and archaeological reconstruction. The approach combined the fragment intact facet properties with fractured boundary curves, employing a new 'fragment topology' representation to capture fractured facet features and their relationships with intact facets. Additionally, a 'ratio of curvature' feature was proposed for fragment facet segmentation and classification. Experiments demonstrated that the use of fragment topology simplified potential match searches and improved the reassembly process, avoiding the complexity and increased computational cost associated with the use of numerous specialized algorithms in each step in the previous methods. The 'ratio of curvature' feature significantly improved facet classification and segmentation, achieving 98.6% recall score when tested on Nidaros Cathedral dataset which outperformed state-of-the-art methods (Mavridis et al. 2015 and Q.-X. Huang et al. 2006) which scored 56% and 50% of recall. This work has been presented as a full paper at ICPRS-19 and published in IET Conference Proceedings (Alzaid and Dogramadzi 2019).

7.3 Future Directions

Although the proposed methods are intended for THR image analysis, these methods can be extended in many directions.

While the current research focused on the development and evaluation of the automated model for segmentation of implant, detection of the Gruen landmarks and diagnosis of PFF, a comparative study involving a diverse group of human experts, including radiologists, surgeons, medical students, and other clinical specialists, is indeed a valuable pathway for future research. Conducting such a comparative study could provide a comprehensive understanding of the performance of the developed system across different levels of clinical expertise and show its strengths and limitations. This could help in assessing the model's clinical relevance and

understanding where it excels or faces challenges. Additionally, such a study could contribute to the validation of the model in a clinical context, ensuring that it aligns with human expertise and can be effectively integrated into the existing diagnostic workflow.

The developed method for the identification of Gruen landmarks and segmentation of the implant component can be extended to compute and analyse many other THR complications such as implant dislocation, subsidence and implant loosening or defining a wide range of implant designs. In addition, this method represents a pioneering approach in the domain of detecting Gruen landmarks, presenting the initial set of results achieved in this domain. As a future direction, exploring more advanced CNN blocks such as Squeeze-and-Excitation Block, Multi-Head Attention and Hypercolumn Blocks with the goal of enhancing these outcomes could be considered. Enhancing the accuracy of landmark detection can significantly contribute to improving the overall PFF diagnosis process, leading to more precise classification and better treatment planning for patients.

Furthermore, the PFF diagnosis approach opens the possibility for its extension to other orthopaedic implant joints, such as the shoulder, knee, or ankle. The outcomes achieved in diagnosing PFFs using the developed framework provide a foundation for adapting and applying this methodology to diverse joint contexts. For instance, similarities between shoulder and hip implants could be evaluated to determine if the segmentation challenges and requirements for the two types of implants align closely. Also, the THR dataset can be adapted to train the segmentation model for shoulder implant and the diagnosis network could be fine-tuned according to the shoulder implant requirements.

As demonstrated in Chapter 4 and Chapter 5 that the integration of shape prior knowledge and clinical interpretation medical knowledge into DL model has significantly contributed to the enhancement of results, effectively addressing challenges arising from dataset limitations. Exploring various dimensions of medical knowledge such as implant type knowledge, surgical approaches and follow-up radiographs, could further improve the effectiveness of the THR analysis approach.

For the reassembling fractured objects method a new representation is defined to minimize the search for potential matching which consisted of multiple features such as the boundary curve of the fracture region, the future direction will include the expansion of the fragment representation to involve additional properties that accurately describe the intact surface. Furthermore, apply

and evaluate the method to more complex and noisy fragments such as 3D models of fractured bones.

References

Abdel, M. P., Watts, C. D., Houdek, M. T., Lewallen, D. G., and Berry, D. J. (2016). "Epidemiology of periprosthetic fracture of the femur in 32 644 primary total hip arthroplasties: A 40-year experience". In: *Bone and Joint Journal* 98B.4, pp. 461–467. ISSN: 20494408. DOI: 10.1302/0301-620X.98B4.37201.

Alom, Md Zahangir, Hasan, Mahmudul, Yakopcic, Chris, Taha, Tarek M, and Asari, Vijayan K (2018). "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation". In: *arXiv preprint arXiv:1802.06955*.

Alzaid, Asma and Dogramadzi, Sanja (2019). "Reassembly of fractured object using fragment topology". In.

Alzaid, Asma, Lineham, Beth, Dogramadzi, Sanja, Pandit, Hemant, and Xie, Sheng Quan (2023). "Simultaneous Hip Implant Segmentation and Gruen Landmarks Detection". In: *Submitted to IEEE Journal of Biomedical Health Informatics*.

Alzaid, Asma, Wignall, Alice, Dogramadzi, Sanja, Pandit, Hemant, and Xie, Sheng Quan (2022). "Automatic detection and classification of peri-prosthetic femur fracture". In: *International Journal of Computer Assisted Radiology and Surgery* 17.4, pp. 649–660.

Ambellan, Felix, Tack, Alexander, Ehlke, Moritz, and Zachow, Stefan (2019). "Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative". In: *Medical image analysis* 52, pp. 109–118.

Amigoni, Francesco, Gazzani, Stefano, and Podico, Simone (2003). "A method for reassembling fragments in image reconstruction". In: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. Vol. 3. IEEE, pp. III–581.

Badgeley, Marcus A., Zech, John R., Oakden-Rayner, Luke, Glicksberg, Benjamin S., Liu, Manway, Gale, William, McConnell, Michael V., Percha, Bethany, Snyder, Thomas M., and Dudley, Joel T. (2019). "Deep learning predicts hip fracture using confounding patient and healthcare variables". In: *npj Digital Medicine* 2.1, p. 31. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0105-1. eprint: 1811.03695.

Banaszkiewicz, Paul A (2014). ""Modes of failure" of cemented stem-type femoral components: a radiographic analysis of loosening". In: *Classic Papers in Orthopaedics*. Springer, pp. 35–38.

Bandyopadhyay, O., Biswas, A., and Bhattacharya, B. B. (Oct. 2016). "Classification of long-bone fractures based on digital-geometric analysis of X-ray images". In: *Pattern Recognition and Image Analysis* 26.4, pp. 742–757. ISSN: 15556212. DOI: 10.1134/S1054661816040027. URL: https://link.springer.com/article/10.1134/S1054661816040027.

Barker, Timothy M and Donnelly, William J (2003). "Automated image analysis technique for measurement of femoral component subsidence in total hip joint replacement". In: *Medical engineering & physics* 25.2, pp. 91–97.

Bayram, Fatih and Çakirotlu, Murat (2016). "DIFFRACT: DIaphyseal Femur FRActure Classifier SysTem". In: *Biocybernetics and Biomedical Engineering* 36.1, pp. 157–171. ISSN: 02085216. DOI: 10.1016/j.bbe.2015.10.003.

Bookstein, Fred L. (1989). "Principal warps: Thin-plate splines and the decomposition of deformations". In: *IEEE Transactions on pattern analysis and machine intelligence* 11.6, pp. 567–585.

Borjali, Alireza, Chen, Antonia F, Bedair, Hany S, Melnic, Christopher M, Muratoglu, Orhun K, Morid, Mohammad A, and Varadarajan, Kartik M (2021). "Comparing the performance of a deep convolutional neural network with orthopedic surgeons on the identification of total hip prosthesis design from plain radiographs". In: *Medical Physics* 48.5, pp. 2327–2336.

Borjali, Alireza, Chen, Antonia F, Muratoglu, Orhun K, Morid, Mohammad A, and Varadarajan, Kartik M (2019). "Detecting mechanical loosening of total hip replacement implant from plain radiograph using deep convolutional neural network". In: *arXiv preprint arXiv:1912.00943*.

Brusini, Irene, Lindberg, Olof, Muehlboeck, J-Sebastian, Smedby, Örjan, Westman, Eric, and Wang, Chunliang (2020). "Shape information improves the cross-cohort performance of deep learning-based segmentation of the hippocampus". In: *Frontiers in neuroscience* 14, p. 15.

Çallı, Erdi, Sogancioglu, Ecem, Ginneken, Bram van, Leeuwen, Kicky G van, and Murphy, Keelin (2021). "Deep learning for chest X-ray analysis: A survey". In: *Medical Image Analysis* 72, p. 102125.

Capone, Antonio, Congia, Stefano, Civinini, Roberto, and Marongiu, Giuseppe (2017). "Periprosthetic fractures: Epidemiology and current treatment". In: *Clinical Cases in Mineral and Bone Metabolism* 14.2, pp. 189–196. ISSN: 19713266. DOI: 10.11138/ccmbm/2017.14.1.189.

Chai, Hum Yan, Wee, Lai Khin, Swee, Tan Tian, and Hussain, Sheikh (2011). "Gray-level co-occurrence matrix bone fracture detection". In: *WSEAS Transactions on Systems* 10.1, pp. 7–16. ISSN: 11092777. URL: http://www.fkbsk.utm.my/.

Chen, Chen, Biffi, Carlo, Tarroni, Giacomo, Petersen, Steffen, Bai, Wenjia, and Rueckert, Daniel (2019). "Learning shape priors for robust cardiac MR segmentation from multi-view images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 523–531.

Chen, Haomin, Wang, Yirui, Zheng, Kang, Li, Weijian, Chang, Chi Tung, Harrison, Adam P., Xiao, Jing, Hager, Gregory D., Lu, Le, Liao, Chien Hung, and Miao, Shun (Aug. 2020). "Anatomy-Aware Siamese Network: Exploiting Semantic Asymmetry for Accurate Pelvic Fracture Detection in X-Ray Images". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12368 LNCS. Springer Science and Business Media Deutschland GmbH, pp. 239–255. ISBN: 9783030585914. DOI: 10.1007/978-3-030-58592-1_15. arXiv: 2007.01464. URL: https://doi.org/10.1007/978-3-030-58592-1%7B%5C_%7D15.

Cheng, Chi-Tung, Ho, Tsung-Ying, Lee, Tao-Yi, Chang, Chih-Chen, Chou, Ching-Cheng, Chen, Chih-Chi, Chung, I-Fang, and Liao, Chien-Hung (2019). "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs". In: *European radiology* 29.10, pp. 5469–5477. DOI: https://doi.org/10.1007/s00330-019-06167-y.

Chowdhury, Muhammad EH, Rahman, Tawsifur, Khandakar, Amith, Mazhar, Rashid, Kadir, Muhammad Abdul, Mahbub, Zaid Bin, Islam, Khandakar Reajul, Khan, Muhammad Salman, Iqbal, Atif, Al Emadi, Nasser, et al. (2020). "Can AI help in screening viral and COVID-19 pneumonia?" In: *Ieee Access* 8, pp. 132665–132676.

Chung, Seok Won, Han, Seung Seog, Lee, Ji Whan, Oh, Kyung-Soo, Kim, Na Ra, Yoon, Jong Pil, Kim, Joon Yub, Moon, Sung Hoon, Kwon, Jieun, Lee, Hyo-Jin, et al. (2018). "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm". In: *Acta orthopaedica* 89.4, pp. 468–473.

Cootes, T.F., Taylor, C.J., Cooper, D.H., and Graham, J. (1995). "Active Shape Models-Their Training and Application". In: *Computer Vision and Image Understanding* 61.1, pp. 38–59. ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1995.1004. URL: https://www.sciencedirect.com/science/article/pii/S1077314285710041.

Debelee, Taye Girma, Schwenker, Friedhelm, Ibenthal, Achim, and Yohannes, Dereje (Mar. 2020). "Survey of deep learning in breast cancer image analysis". In: *Evolving Systems* 11.1, pp. 143–163. ISSN: 18686486. DOI: 10.1007/S12530-019-09297-2.

Dellepiane, M, Niccolucci, F, Serna, S Pena, Rushmeier, H, Van Gool, L, et al. (2011). "Reassembling thin artifacts of unknown geometry". In.

Dhiman, Gaurav, Juneja, Sapna, Viriyasitavat, Wattana, Mohafez, Hamidreza, Hadizadeh, Maryam, Islam, Mohammad Aminul, El Bayoumy, Ibrahim, and Gulati, Kamal (2022). "A novel machine-learning-based hybrid CNN model for tumor identification in medical image processing". In: *Sustainability* 14.3, p. 1447.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain,

et al. (2020). “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929*.

Downing, Martin Richard, Undrill, Peter Edward, Ashcroft, Paddy, Hukins, DWL, and Hutchison, JD (1997). “Automated femoral measurement in total hip replacement radiographs”. In.

Duan, Jinming, Bello, Ghalib, Schlemper, Jo, Bai, Wenjia, Dawes, Timothy JW, Biffi, Carlo, Marvao, Antonio de, Doumoud, Georgia, O’Regan, Declan P, and Rueckert, Daniel (2019). “Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach”. In: *IEEE transactions on medical imaging* 38.9, pp. 2151–2164.

Ebrahimighahnavieh, Mr Amir, Luo, Suhuai, and Chiong, Raymond (Apr. 2020). “Deep learning to detect Alzheimer’s disease from neuroimaging: A systematic literature review”. In: *Computer Methods and Programs in Biomedicine* 187, p. 105242. ISSN: 0169-2607. DOI: 10.1016/J.CMPB.2019.105242.

Ebramzadeh, Edward, Normand, Patricia L, Sangiorgio, Sophia N, Llinás, Adolfo, Gruen, Thomas A, McKellop, Harry A, and Sarmiento, Augusto (2003). “Long-term radiographic changes in cemented total hip arthroplasty with six designs of femoral components”. In: *Biomaterials* 24.19, pp. 3351–3363.

ENGH, CHARLES A, Massin, Philippe, and SUTHERS, KATHLEEN E (1990). “Roentgenographic assessment of the biologic fixation of porous-surfaced femoral components.” In: *Clinical Orthopaedics and Related Research (1976-2007)* 257, pp. 107–128.

Fischler, Martin A and Bolles, Robert C (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24.6, pp. 381–395.

Florea, Laura, Florea, Corneliu, Vertan, Constantin, and Sultana, Alina (2011). “Automatic tools for diagnosis support of total hip replacement follow-up”. In: *Advances in Electrical and Computer Engineering* 11.4, pp. 55–62.

Florea, Laura and Vertan, Constantin (2009). “Automatic Hip Prosthesis Fit Estimation By Cooperative X-ray Image Segmentation”. In: 71.

Foran, Jared R H and Fischer, Stuart James (2020). *Total Hip Replacement*. URL: `https://orthoinfo.aaos.org/en/treatment/total-hip-replacement/`.

Fürnstahl, Philipp, Székely, Gábor, Gerber, Christian, Hodler, Jürg, Snedeker, Jess Gerrit, and Harders, Matthias (2012). "Computer assisted reconstruction of complex proximal humerus fractures for preoperative planning". In: *Medical image analysis* 16.3, pp. 704–720.

Gama Leitão, Helena Cristina da and Stolfi, Jorge (2002). "A multiscale method for the reassembly of two-dimensional fragmented objects". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.9, pp. 1239–1251.

Gan, Hong-Seng, Ramlee, Muhammad Hanif, Wahab, Asnida Abdul, Lee, Yeng-Seng, and Shimizu, Akinobu (2021). "From classical to deep learning: review on cartilage and bone segmentation techniques in knee osteoarthritis research". In: *Artificial Intelligence Review* 54.4, pp. 2445–2494.

Gao, Jun, Jiang, Qian, Zhou, Bo, and Chen, Daozheng (2019). "Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview". In: *Mathematical Biosciences and Engineering* 16.6, pp. 6536–6561.

Goldberger, Ary L, Amaral, Luis AN, Glass, Leon, Hausdorff, Jeffrey M, Ivanov, Plamen Ch, Mark, Roger G, Mietus, Joseph E, Moody, George B, Peng, Chung-Kang, and Stanley, H Eugene (2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23, e215–e220.

Gong, Zibo, Fu, Yonghui, He, Ming, and Fu, Xinzhe (2022). "Automated identification of hip arthroplasty implants using artificial intelligence". In: *Scientific Reports* 12.1, p. 12179.

Gonzalez Rafael, C (2008). *Digital Image Processing, /Rafael C. Gonzalez, Richard E. Woods.*

GRUEN, THOMAS A. M.S., MCNEICE, GREGORY M., and AMSTUTZ, HARLAN C. M.D. (1979). "'Modes of Failure' of Cemented Stem-type Femoral Components: A Radiographic Analysis of Loosening." In: *Clinical Orthopaedics and Related Research*, pp. 17–27.

Gu, Zaiwang, Cheng, Jun, Fu, Huazhu, Zhou, Kang, Hao, Huaying, Zhao, Yitian, Zhang, Tianyang, Gao, Shenghua, and Liu, Jiang (2019). "Ce-net: Context encoder network for 2d

medical image segmentation". In: *IEEE transactions on medical imaging* 38.10, pp. 2281–2292.

Gupta, Vikash, Demirer, Mutlu, Bigelow, Matthew, Yu, Sarah M., Yu, Joseph S., Prevedello, Luciano M., White, Richard D., and Erdal, Barbaros S. (Apr. 2020). "Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs". In: *Proceedings - International Symposium on Biomedical Imaging* 2020-April, pp. 1526–1529. ISSN: 19458452. DOI: 10.1109/ISBI45749.2020.9098436.

Halder, Amitava, Dey, Debangshu, and Sadhu, Anup K (2020). "Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review". In: *Journal of digital imaging* 33.3, pp. 655–677.

He, Joshua Congfu, Leow, Wee Kheng, and Howe, Tet Sen (2007). "Hierarchical classifiers for detection of fractures in X-ray images". In: *International Conference on Computer Analysis of Images and Patterns*. Springer, pp. 962–969. ISBN: 9783540742715. DOI: 10.1007/978-3-540-74272-2-119.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hmeidi, Ismail, Al-Ayyoub, Mahmoud, Rababah, Haya, and Khatatbeh, Zakaria (2013). "Detecting Hand Bone Fractures in X-Ray Images". In: *JMPT* 4.3, pp. 155–168. DOI: 10.13140/rg.2.1.2645.8327.

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

Huang, Gao, Liu, Zhuang, Maaten, Laurens van der, and Weinberger, Kilian Q. (July 2017). "Densely Connected Convolutional Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, Qi-Xing, Flöry, Simon, Gelfand, Natasha, Hofer, Michael, and Pottmann, Helmut (July 2006). "Reassembling Fractured Objects by Geometric Matching". In: *ACM Trans. Graph.* 25.3, pp. 569–578. ISSN: 0730-0301. DOI: 10.1145/1141911.1141925. URL: http://doi.acm.org/10.1145/1141911.1141925.

Jia, Li-qin, Peng, Cheng-zhang, Liu, Hong-min, and Wang, Zhi-heng (2011). "A fast randomized circle detection algorithm". In: *2011 4th International Congress on Image and Signal Processing*. Vol. 2. IEEE, pp. 820–823.

Jiménez-Sánchez, Amelia, Kazi, Anees, Albarqouni, Shadi, Kirchhoff, Chlodwig, Biberthaler, Peter, Navab, Nassir, Kirchhoff, Sonja, and Mateus, Diana (2020). "Precise proximal femur fracture classification for interactive training and surgical planning". In: *International Journal of Computer Assisted Radiology and Surgery* 15.5, pp. 847–857. ISSN: 18616429. DOI: 10.1007/s11548-020-02150-x. eprint: 1902.01338.

Jiménez-Sánchez, Amelia, Kazi, Anees, Albarqouni, Shadi, Kirchhoff, Sonja, Sträter, Alexandra, Biberthaler, Peter, Mateus, Diana, and Navab, Nassir (2018). "Weakly-Supervised Localization and Classification of Proximal Femur Fractures". In: pp. 1–7. arXiv: 1809.10692. URL: http://arxiv.org/abs/1809.10692.

Jiménez-Sánchez, Amelia, Mateus, Diana, Kirchhoff, Sonja, Kirchhoff, Chlodwig, Biberthaler, Peter, Navab, Nassir, González Ballester, Miguel A, and Piella, Gemma (2019). "Medical-based Deep Curriculum Learning for Improved Fracture Classification". In: DOI: 10.1007/978-3-030-32226-7. URL: https://doi.org/10.1007/978-3-030-32226-7_77.

Johnson, Justin M and Khoshgoftaar, Taghi M (2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6.1, pp. 1–54.

Joshi, Deepa and Singh, Thipendra P (2020). "A survey of fracture detection techniques in bone X-ray images". In: *Artificial Intelligence Review* 53.6, pp. 4475–4517.

Kang, Yang-Jae, Yoo, Jun-Il, Cha, Yong-Han, Park, Chan H, and Kim, Jung-Taek (2020). "Machine learning–based identification of hip arthroplasty designs". In: *Journal of orthopaedic translation* 21, pp. 13–17.

Karimi, Davood, Samei, Golnoosh, Kesch, Claudia, Nir, Guy, and Salcudean, Septimiu E (2018). "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models". In: *International journal of computer assisted radiology and surgery* 13.8, pp. 1211–1219.

Karnuta, Jaret M, Haeberle, Heather S, Luu, Bryan C, Roth, Alexander L, Molloy, Robert M, Nystrom, Lukas M, Piuzzi, Nicolas S, Schaffer, Jonathan L, Chen, Antonia F, Iorio, Richard, et al. (2021). "Artificial intelligence to identify arthroplasty implants from radiographs of the hip". In: *The Journal of arthroplasty* 36.7, S290–S294.

Kavur, A Emre, Gezer, Naciye Sinem, Barış, Mustafa, Şahin, Yusuf, Özkan, Savaş, Baydar, Bora, Yüksel, Ulaş, Kılıkçıer, Çağlar, Olut, Şahin, Akar, Gözde Bozdağı, et al. (2020). "Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors". In: *Diagnostic and Interventional Radiology* 26.1, p. 11.

Khosravi, Bardia, Rouzrokh, Pouria, Maradit Kremers, Hilal, Larson, Dirk R, Johnson, Quinn J, Faghani, Shahriar, Kremers, Walter K, Erickson, Bradley J, Sierra, Rafael J, Taunton, Michael J, et al. (2022). "Patient-specific hip arthroplasty dislocation risk calculator: an explainable multimodal machine learning–based approach". In: *Radiology: Artificial Intelligence* 4.6, e220067.

Kim, D. H. and MacKinnon, T. (May 2018). "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks". In: *Clinical Radiology* 73.5, pp. 439–445. ISSN: 1365229X. DOI: 10.1016/j.crad.2017.11.015.

Kitamura, Gene, Chung, Chul Y., and Moore, Barry E. (2019). "Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation". In: *Journal of Digital Imaging* 32.4, pp. 672–677. ISSN: 1618727X. DOI: 10.1007/s10278-018-0167-7.

Kong, Weixin and Kimia, Benjamin B (2001). "On solving 2D and 3D puzzles using curve matching". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on.* Vol. 2. IEEE, pp. II–II.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25.

Krogue, Justin D, Cheng, Kaiyang, Hwang, Kevin M, Toogood, Paul, Meinberg, Eric G., Geiger, Erik J, Zaid, Musa, McGill, Kevin C., Patel, Rina, Sohn, Jae Ho, Wright, Alexandra, Darger, Bryan F, Padrez, Kevin A, Ozhinsky, Eugene, Majumdar, Sharmila, and Pedoia, Valentina (2019). *Automatic hip fracture identification and functional subclassification with deep learning*. DOI: 10.1148/ryai.2020190023. arXiv: 1909.06326.

Kwong, Timothy and Mazaheri, Samaneh (2021). "A survey on deep learning approaches for breast cancer diagnosis". In: *arXiv preprint arXiv:2109.08853*.

Lai, Jiing-Yih and Chen, Kuo-Jen (2007). "Localization of parts with irregular shape for CMM inspection". In: *The International Journal of Advanced Manufacturing Technology* 32.11, pp. 1188–1200.

Larrazabal, Agostina J, Martinez, Cesar, and Ferrante, Enzo (2019). "Anatomical priors for image segmentation via post-processing with denoising autoencoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 585–593.

Learmonth, Ian D, Young, Claire, and Rorabeck, Cecil (2007). "The operation of the century: total hip replacement". In: *The Lancet* 370.9597, pp. 1508–1519. DOI: 10.1016/S0140-6736(07)60457-7.

Lee, Pei-Yuan, Lai, Jiing-Yih, Yu, Shou-An, Huang, Chung-Yi, Hu, Yu-Sheng, and Feng, Chien-Lin (2014). "Computer-assisted fracture reduction and fixation simulation for pelvic fractures". In: *Journa of Medical and Biological Engineering* 34, pp. 368–376.

Lee, Shanjean, Kagan, Ryland, Wang, Lian, and Doung, Yee Cheen (2019). "Reliability and Validity of the Vancouver Classification in Periprosthetic Fractures Around Cementless Femoral Stems". In: *Journal of Arthroplasty* 34.7, S277–S281. DOI: 10.1016/j.arth.2019.02.062.

Li, Liu, Xu, Mai, Wang, Xiaofei, Jiang, Lai, and Liu, Hanruo (2019). "Attention based glaucoma detection: a large-scale database and CNN model". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10571–10580.

Li, Xin, Qin, Genggeng, He, Qiang, Sun, Lei, Zeng, Hui, He, Zilong, Chen, Weiguo, Zhen, Xin, and Zhou, Linghong (2020). "Digital breast tomosynthesis versus digital mammography: integration of image modalities enhances deep learning-based breast mass classification". In: *European radiology* 30, pp. 778–788.

Li, Yuanwei, Ho, Chin Pang, Toulemonde, Matthieu, Chahal, Navtej, Senior, Roxy, and Tang, Meng-Xing (2017). "Fully automatic myocardial segmentation of contrast echocardiography sequence using random forests guided by shape model". In: *IEEE transactions on medical imaging* 37.5, pp. 1081–1091.

Lin, Tsung Yi, Goyal, Priya, Girshick, Ross, He, Kaiming, and Dollár, Piotr (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Lindsey, Robert, Daluiski, Aaron, Chopra, Sumit, Lachapelle, Alexander, Mozer, Michael, Sicular, Serge, Hanel, Douglas, Gardner, Michael, Gupta, Anurag, Hotchkiss, Robert, et al. (2018). "Deep neural network improves fracture detection by clinicians". In: *Proceedings of the National Academy of Sciences* 115.45, pp. 11591–11596.

Liu, Quande, Yu, Lequan, Luo, Luyang, Dou, Qi, and Heng, Pheng Ann (2020). "Semi-supervised medical image classification with relation-driven self-ensembling model". In: *IEEE transactions on medical imaging* 39.11, pp. 3429–3440.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, and Guo, Baining (2021). "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.

Loppini, Mattia, Gambaro, Francesco Manlio, Chiappetta, Katia, Grappiolo, Guido, Bianchi, Anna Maria, and Corino, Valentina DA (2022). "Automatic Identification of Failure in Hip Replacement: An Artificial Intelligence Approach". In: *Bioengineering* 9.7, p. 288.

Lotfy, Mayar, Shubair, Raed M., Navab, Nassir, and Albarqouni, Shadi (Nov. 2019). "Investigation of Focal Loss in Deep Learning Models for Femur Fractures Classification". In: *2019 International Conference on Electrical and Computing Technologies and Applications, ICECTA 2019*. DOI: 10.1109/ICECTA48151.2019.8959770.

Luo, Jun, Kitamura, Gene, Arefan, Dooman, Doganay, Emine, Panigrahy, Ashok, and Wu, Shandong (n.d.). "Knowledge-Guided Multiview Deep Curriculum Learning for Elbow Fracture Classification". In: (). arXiv: 2110.10383v1. URL: https://github.com/ljaiverson/multiview-curriculum..

Maggs, Joanna L., Swanton, Eric, Whitehouse, Sarah L., Howell, Jonathan R., Timperley, A. John, Hubble, Matthew J. W., and Wilson, Matt J. (2021). "B2 or not B2? That is the question: a review of periprosthetic fractures around cemented taper-slip femoral components". In: *The Bone and Joint Journal* 103-B.1, pp. 71–78. DOI: 10.1302/0301-620X.103B1.BJJ-2020-0163.R1. URL: https://doi.org/10.1302/0301-620X.103B1.BJJ-2020-0163.R1.

Maghdid, Halgurd S, Asaad, Aras T, Ghafoor, Kayhan Zrar, Sadiq, Ali Safaa, Mirjalili, Seyedali, and Khan, Muhammad Khurram (2021). "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms". In: *Multimodal image exploitation and learning 2021*. Vol. 11734. SPIE, pp. 99–110.

Mahendran, S K and Santhosh Baboo, S (2011). "An Enhanced Tibia Fracture Detection Tool Using Image Processing and Classification Fusion Techniques in X-Ray Images ". In: *Global Journal of Computer Science and Technology* 11. ISSN: 0975-4350.

Marshall, Richard A., Weaver, Michael J., Sodickson, Aaron, and Khurana, Bharti (2017). "Periprosthetic femoral fractures in the emergency department: What the orthopedic surgeon wants to know". In: *Radiographics* 37.4, pp. 1202–1217. ISSN: 15271323. DOI: 10.1148/rg.2017160127.

Mavridis, P, Andreadis, A, and Papaioannou, G (2015). "Fractured Object Reassembly via Robust Surface Registration". In: *Eurographics*. URL: https://pdfs.semanticscholar.org/e3ed/6c2a200dc8e4f00c38813cda202a949821cb.pdf%20http://www.presious.eu/sites/default/files/EG15%7B%5C_%7DReassembly%7B%5C_%7Dfinal.pdf.

McBride, Jonah C and Kimia, Benjamin B (2003). "Archaeological fragment reconstruction using curve-matching". In: *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*. Vol. 1. IEEE, pp. 3–3.

McBride, Tim J and Prakash, Divya (Feb. 2011). "How to read a postoperative total hip replacement radiograph". In: *Postgraduate Medical Journal* 87.1024, pp. 101–109. ISSN: 0032-5473. DOI: 10.1136/PGMJ.2009.095620. URL: https://pmj.bmj.com/content/87/1024/101%20https://pmj.bmj.com/content/87/1024/101.abstract.

Medley, Daniela O, Santiago, Carlos, and Nascimento, Jacinto C (2019). "Deep active shape model for robust object fitting". In: *IEEE Transactions on Image Processing* 29, pp. 2380–2394.

Meedeniya, Dulani, Kumarasinghe, Hashara, Kolonne, Shammi, Fernando, Chamodi, De la Torre Díez, Isabel, and Marques, Goncalo (2022). "Chest X-ray analysis empowered with deep learning: A systematic review". In: *Applied Soft Computing*, p. 109319.

Mehnert, Andrew and Jackway, Paul (Oct. 1997). "An improved seeded region growing algorithm". In: *Pattern Recognition Letters* 18.10, pp. 1065–1071. ISSN: 0167-8655. URL: https://www.sciencedirect.com/science/article/pii/S0167865597001311.

Mellado, Nicolas, Reuter, Patrick, and Schlick, Christophe (2010). "Semi-automatic geometry-driven reassembly of fractured archeological objects". In: *VAST 2010: The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, p. 00.

Miao, Yu, Zhao, Peng-Fei, Tang, Xiong-Feng, Li, Yu-Qin, Zhang, Li-Yuan, Shi, Wei-Li, Zhang, Ke, Yang, Hua-Min, and Liu, Jian-Hua (2019). "A method for detecting femur fracture based on SK-DenseNet". In: *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing*. Vol. 7. ACM, pp. 1–7. ISBN: 9781450372022. DOI: 10.1145/3358331.3358402.

Mirikharaji, Zahra and Hamarneh, Ghassan (2018). "Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos

Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, pp. 737–745. ISBN: 978-3-030-00937-3.

Muscato, Federico, Corti, Anna, Gambaro, Francesco Manlio, Chiappetta, Katia, Loppini, Mattia, and Corino, Valentina DA (2023). "Combining deep learning and machine learning for the automatic identification of hip prosthesis failure: Development, validation and explainability analysis". In: *International Journal of Medical Informatics* 176, p. 105095.

Naser, Mohamed A and Deen, M Jamal (2020). "Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images". In: *Computers in biology and medicine* 121, p. 103758.

Nguyen, Duy HM, Nguyen, Duy M, Mai, Truong TN, Nguyen, Thu, Tran, Khanh T, Nguyen, Anh Triet, Pham, Bao T, and Nguyen, Binh T (2022). "ASMCNN: An efficient brain extraction using active shape model and convolutional neural networks". In: *Information Sciences* 591, pp. 25–48.

Niblack, Wayne (1985). *An introduction to digital image processing.* Strandberg Publishing Company.

OECD (2019). *Hip and knee replacement*, pp. 198, 199. DOI: https://doi.org/https://doi.org/10.1787/2fc83b9a-en. URL: https://www.oecd-ilibrary.org/content/component/2fc83b9a-en.

Oktay, Ozan, Schlemper, Jo, Folgoc, Loic Le, Lee, Matthew, Heinrich, Mattias, Misawa, Kazunari, Mori, Kensaku, McDonagh, Steven, Hammerla, Nils Y, Kainz, Bernhard, et al. (2018). "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999*.

Olczak, Jakub, Fahlberg, Niklas, Maki, Atsuto, Razavian, Ali Sharif, Jilert, Anthony, Stark, André, Sköldenberg, Olof, and Gordon, Max (2017). "Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures?" In: *Acta Orthopaedica* 88.6, pp. 581–586. ISSN: 17453682. DOI: 10.1080/17453674.2017.1344459.

Oprea, Alina and Vertan, Constantin (2007). "A quantitative evaluation of the hip prosthesis segmentation quality in X-ray images". In: *2007 International Symposium on Signals, Circuits and Systems*. Vol. 1. IEEE, pp. 1–4.

Painchaud, Nathan, Skandarani, Youssef, Judge, Thierry, Bernard, Olivier, Lalande, Alain, and Jodoin, Pierre-Marc (2020). "Cardiac segmentation with strong anatomical guarantees". In: *IEEE transactions on medical imaging* 39.11, pp. 3703–3713.

Papaioannou, Georgios, Karabassi, E-A, and Theoharis, Theoharis (2002). "Reconstruction of three-dimensional objects through matching of their parts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.1, pp. 114–124.

Papaioannou, Georgios and Karabassi, Evaggelia-Aggeliki (2003). "On the automatic assemblage of arbitrary broken solid artefacts". In: *Image and Vision Computing* 21.5, pp. 401–412.

Patel, Ravi, Thong, Elizabeth HE, Batta, Vineet, Bharath, Anil Anthony, Francis, Darrel, and Howard, James (2021). "Automated Identification of Orthopedic Implants on Radiographs Using Deep Learning". In: *Radiology: Artificial Intelligence* 3.4, e200183.

Powell-Bowns, Matilda F. R., Oag, Erlend, Ng, Nathan, Pandit, Hemant, Moran, Matthew, Patton, James T., Clement, Nick D., and Scott, Chloe E. H. (2021). "Vancouver B periprosthetic fractures involving the Exeter cemented stem". In: *The Bone and Joint Journal* 103-B.2, pp. 309–320. DOI: 10.1302/0301-620X.103B2.BJJ-2020-0695.R1. URL: https://doi.org/10.1302/0301-620X.103B2.BJJ-2020-0695.R1.

Qin, Chunxia, Tu, Puxun, Chen, Xiaojun, and Troccaz, Jocelyne (2022). "A novel registration-based algorithm for prostate segmentation via the combination of SSM and CNN". In: *Medical Physics*.

Qin, Xuebin, Zhang, Zichen, Huang, Chenyang, Dehghan, Masood, Zaiane, Osmar R, and Jagersand, Martin (2020). "U2-Net: Going deeper with nested U-structure for salient object detection". In: *Pattern recognition* 106, p. 107404.

Rahman, Tawsifur, Khandakar, Amith, Islam, Khandaker Reajul, Soliman, Md Mohiuddin, Islam, Mohammad Tariqul, Elsayed, Ahmed, Qiblawey, Yazan, Mahmud, Sakib, Rahman,

Ashiqur, Musharavati, Farayi, et al. (2022). "HipXNet: Deep Learning Approaches to Detect Aseptic Loos-Ening of Hip Implants Using X-Ray Images". In: *IEEE Access* 10, pp. 53359–53373.

Rajpurkar, Pranav, Irvin, Jeremy, Bagul, Aarti, Ding, Daisy, Duan, Tony, Mehta, Hershel, Yang, Brandon, Zhu, Kaylie, Laird, Dillon, Ball, Robyn L, Langlotz, Curtis, Shpanskaya, Katie, Lungren, Matthew P, and Ng, Andrew Y (2017). *MURA: Large dataset for abnormality detection in musculoskeletal radiographs*. arXiv: 1712.06957. URL: http://stanfordmlgroup.github.io/competitions/mura.%20http://arxiv.org/abs/1712.06957.

Rajpurkar, Pranav, Irvin, Jeremy, Zhu, Kaylie, Yang, Brandon, Mehta, Hershel, Duan, Tony, Ding, Daisy, Bagul, Aarti, Langlotz, Curtis, Shpanskaya, Katie, et al. (2017). "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225*.

Ramavath, Ashoklal, Lamb, Jonathan N, Palan, Jeya, Pandit, Hemant G, and Jain, Sameer (2020). "Postoperative periprosthetic femoral fracture around total hip replacements: current concepts and clinical outcomes". In: *EFORT Open Reviews* 5.9, pp. 558–567. DOI: 10.1302/2058-5241.5.200003.

Ranjbarzadeh, Ramin, Dorosti, Shadi, Ghoushchi, Saeid Jafarzadeh, Caputo, Annalina, Tirkolaee, Erfan Babaee, Ali, Sadia Samar, Arshadi, Zahra, and Bendechache, Malika (2022). "Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods". In: *Computers in Biology and Medicine*, p. 106443.

Ren, Shaoqing, He, Kaiming, Girshick, Ross, and Sun, Jian (2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6, pp. 1137–1149. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2577031. eprint: 1506.01497.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, pp. 234–241.

Roth, Holger R., Wang, Yinong, Yao, Jianhua, Lu, Le, Burns, Joseph E., and Summers, Ronald M. (2016). "Deep convolutional networks for automated detection of posterior-element fractures on spine CT". In: *Medical Imaging 2016: Computer-Aided Diagnosis*. Vol. 9785. SPIE. DOI: 10.1117/12.2217146.

Rouzrokh, Pouria, Ramazanian, Taghi, Wyles, Cody C, Philbrick, Kenneth A, Cai, Jason C, Taunton, Michael J, Kremers, Hilal Maradit, Lewallen, David G, and Erickson, Bradley J (2021). "Deep learning artificial intelligence model for assessment of hip dislocation risk following primary Total hip arthroplasty from postoperative radiographs". In: *The Journal of Arthroplasty* 36.6, pp. 2197–2203.

Rouzrokh, Pouria, Wyles, Cody C, Kurian, Shyam J, Ramazanian, Taghi, Cai, Jason C, Huang, Qiao, Zhang, Kuan, Taunton, Michael J, Maradit Kremers, Hilal, and Erickson, Bradley J (2022). "Deep learning for radiographic measurement of femoral component subsidence following total hip arthroplasty". In: *Radiology: Artificial Intelligence* 4.3, e210206.

Rouzrokh, Pouria, Wyles, Cody C, Philbrick, Kenneth A, Ramazanian, Taghi, Weston, Alexander D, Cai, Jason C, Taunton, Michael J, Lewallen, David G, Berry, Daniel J, Erickson, Bradley J, et al. (2021). "A deep learning tool for automated radiographic measurement of acetabular component inclination and version after total hip arthroplasty". In: *The Journal of arthroplasty* 36.7, pp. 2510–2517.

Sagiroglu, M.S. and Ercil, A. (Aug. 2006). "A Texture Based Matching Approach for Automated Assembly of Puzzles". In: *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, pp. 1036–1041. ISBN: 0-7695-2521-0. DOI: 10.1109/ICPR.2006.184. URL: http://ieeexplore.ieee.org/document/1699703/.

Schock, Justus, Kopaczka, Marcin, Agthe, Benjamin, Huang, Jie, Kruse, Paul, Truhn, Daniel, Conrad, Stefan, Antoch, Gerald, Kuhl, Christiane, Nebelung, Sven, et al. (2020). "A method for semantic knee bone and cartilage segmentation with deep 3D shape fitting using data from the Osteoarthritis Initiative". In: *International Workshop on Shape in Medical Imaging*. Springer, pp. 85–94.

Schwarzkopf, Ran, Oni, Julius K, and Marwin, Scott E (2013). "Total hip arthroplasty periprosthetic femoral fractures: a review of classification and current treatment." In: *Bulletin of the Hospital for Joint Disease (2013)* 71.1, pp. 68–78. ISSN: 2328-5273 (Electronic).

Sebastian, Thomas B., Klein, Philip N., and Kimia, Benjamin B. (2003). "On aligning curves". In: *IEEE transactions on pattern analysis and machine intelligence* 25.1, pp. 116–125.

Shah, Romil F, Bini, Stefano A, Martinez, Alejandro M, Pedoia, Valentina, and Vail, Thomas P (2020). "Incremental inputs improve the automated detection of implant loosening using machine-learning algorithms". In: *The Bone and Joint Journal* 102.6 Supple A, pp. 101–106.

Shao, Feifei, Chen, Long, Shao, Jian, Ji, Wei, Xiao, Shaoning, Ye, Lu, Zhuang, Yueting, and Xiao, Jun (2021). "Deep Learning for Weakly-Supervised Object Detection and Object Localization: A Survey". In: *arXiv preprint arXiv:2105.12694*.

Simonyan, Karen and Zisserman, Andrew (Sept. 2015). "Very deep convolutional networks for large-scale image recognition". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv: 1409.1556. URL: https://arxiv.org/abs/1409.1556.

Skandarani, Youssef, Jodoin, Pierre-Marc, and Lalande, Alain (2023). "Gans for medical image synthesis: An empirical study". In: *Journal of Imaging* 9.3, p. 69.

Son, Tae-geun, Lee, Jusung, Lim, Jeonghun, and Lee, Kunwoo (Oct. 2018). "Reassembly of fractured objects using surface signature". In: *The Visual Computer* 34.10, pp. 1371–1381. ISSN: 0178-2789. DOI: 10.1007/s00371-017-1419-0. URL: http://link.springer.com/10.1007/s00371-017-1419-0.

Stark, MBCG (2018). "Automatic detection and segmentation of shoulder implants in x-ray images". PhD thesis. Master's Thesis, San Francisco State University, San Francisco, CA, USA ...

Suzuki, Satoshi et al. (1985). "Topological structural analysis of digitized binary images by border following". In: *Computer vision, graphics, and image processing* 30.1, pp. 32–46.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June. IEEE Computer Society, pp. 1–9. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298594. arXiv: 1409.4842.

Tabrizi, Pooneh R, Mansoor, Awais, Cerrolaza, Juan J, Jago, James, and Linguraru, Marius George (2018). "Automatic kidney segmentation in 3D pediatric ultrasound images using deep neural networks and weighted fuzzy active shape model". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 1170–1173.

Tajbakhsh, Nima, Jeyaseelan, Laura, Li, Qian, Chiang, Jeffrey N, Wu, Zhihao, and Ding, Xiaowei (2020). "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation". In: *Medical Image Analysis* 63, p. 101693.

Tanzi, Leonardo, Vezzetti, Enrico, Moreno, Rodrigo, and Moos, Sandro (2020). "X-Ray Bone Fracture Classification Using Deep Learning: A Baseline for Designing a Reliable Approach". In: *Applied Sciences* 10.4, p. 1507. ISSN: 2076-3417. DOI: 10.3390/app10041507.

Thomas, Thaddeus P., Anderson, Donald D., Willis, Andrew R., Liu, Pengcheng, Frank, Matthew C., Marsh, J. Lawrence, and Brown, Thomas D. (Mar. 2011). "A computational/experimental platform for investigating three-dimensional puzzle solving of comminuted articular fractures". In: *Computer Methods in Biomechanics and Biomedical Engineering* 14.3, pp. 263–270. ISSN: 1025-5842. URL: http://www.tandfonline.com/doi/abs/10.1080/10255841003762042.

Tilborghs, Sofie, Dresselaers, Tom, Claus, Piet, Bogaert, Jan, and Maes, Frederik (2020). "Shape constrained CNN for cardiac MR segmentation with simultaneous prediction of shape and pose parameters". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, pp. 127–136.

Tombari, Federico, Salti, Samuele, and Di Stefano, Luigi (2010). "Unique signatures of histograms for local surface description". In: *European Conference on Computer Vision*. Springer, pp. 356–369.

Üçoluk, Göktürk and Toroslu, I Hakkı (1999). "Automatic reconstruction of broken 3-D surface objects". In: *Computers & Graphics* 23.4, pp. 573–582.

Umadevi, N. and Geethalakshmi, S. N. (2012). "Multiple classification system for fracture detection in human bone x-ray images". In: *2012 3rd International Conference on Computing, Communication and Networking Technologies, ICCCNT 2012*. DOI: 10.1109/ICCCNT.2012.6395889.

United Kingdom National Joint Registry (2020). *2020 17th Annual Report*. Tech. rep., pp. 1–248. URL: www.njrreports.org.uk..

Urakawa, Takaaki, Tanaka, Yuki, Goto, Shinichi, Matsuzawa, Hitoshi, Watanabe, Kei, and Endo, Naoto (Feb. 2019). "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network". In: *Skeletal Radiology* 48.2, pp. 239–244. ISSN: 14322161. DOI: 10.1007/S00256-018-3016-3/FIGURES/5. URL: https://link.springer.com/article/10.1007/s00256-018-3016-3.

Vanrusselt, Jan, Vansevenant, Milan, Vanderschueren, Geert, and Vanhoenacker, Filip (2015). "Postoperative radiograph of the hip arthroplasty: what the radiologist should know". In: *Insights into imaging* 6, pp. 591–600.

Varma, Maya, Lu, Mandy, Gardner, Rachel, Dunnmon, Jared, Khandwala, Nishith, Rajpurkar, Pranav, Long, Jin, Beaulieu, Christopher, Shpanskaya, Katie, Fei-Fei, Li, Lungren, Matthew P, and Patel, Bhavik N (2019). "Automated abnormality detection in lower extremity radiographs using deep learning". In: *Nature Machine Intelligence* 1.12, pp. 578–583. DOI: 10.1038/s42256-019-0126-0. URL: https://doi.org/10.1038/s42256-019-0126-0.

Vesal, Sulaiman, Gu, Mingxuan, Maier, Andreas, and Ravikumar, Nishant (2020). "Spatio-temporal multi-task learning for cardiac MRI left ventricle quantification". In: *IEEE Journal of Biomedical and Health Informatics* 25.7, pp. 2698–2709.

Wang, Kun, Zhang, Xiaohong, Huang, Sheng, Chen, Feiyu, Zhang, Xiangbo, and Huangfu, Luwen (2020). "Learning to recognize thoracic disease in chest x-rays with knowledge-guided deep zoom neural networks". In: *IEEE Access* 8, pp. 159790–159805.

Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Wang, Yirui, Lu, Le, Cheng, Chi Tung, Jin, Dakai, Harrison, Adam P., Xiao, Jing, Liao, Chien Hung, and Miao, Shun (2019). "Weakly Supervised Universal Fracture Detection in Pelvic X-Rays". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 11769 LNCS. Springer. Springer, pp. 459–467. DOI: 10.1007/978-3-030-32226-7-51.

Willis, Andrew, Anderson, Donald, Thomas, Thad, Brown, Thomas, and Marsh, J Lawrence (2007). "3D reconstruction of highly fragmented bone fractures". In: *Medical imaging*. International Society for Optics and Photonics, 65121P–65121P.

Xie, Xiaozheng, Niu, Jianwei, Liu, Xuefeng, Chen, Zhengsu, Tang, Shaojie, and Yu, Shui (2021). "A survey on incorporating domain knowledge into deep learning for medical image analysis". In: *Medical Image Analysis* 69, p. 101985.

Xing, Fuyong, Xie, Yuanpu, and Yang, Lin (2015). "An automatic learning-based framework for robust nucleus segmentation". In: *IEEE transactions on medical imaging* 35.2, pp. 550–566.

Yahalomi, Erez, Chernofsky, Michael, and Werman, Michael (2019). "Detection of Distal Radius Fractures Trained by a Small Set of X-Ray Images and Faster R-CNN". In: *Advances in Intelligent Systems and Computing* 997, pp. 971–981. ISSN: 21945365. DOI: 10.1007/978-3-030-22871-2_69/COVER. arXiv: 1812.09025. URL: https://link.springer.com/chapter/10.1007/978-3-030-22871-2_69.

Yan, Rui, Ren, Fei, Wang, Zihao, Wang, Lihua, Zhang, Tong, Liu, Yudong, Rao, Xiaosong, Zheng, Chunhou, and Zhang, Fa (2020). "Breast cancer histopathological image classification using a hybrid deep neural network". In: *Methods* 173, pp. 52–60.

Yin, Zhao, Wei, Li, Li, Xin, and Manhein, Mary (2011). "An automatic assembly and completion framework for fragmented skulls". In: *2011 International Conference on Computer Vision*. IEEE, pp. 2532–2539.

Yoon, Sun-Jung, Kim, Tae Hyong, Joo, Su-Bin, and Oh, Seung Eel (2020). "Automatic multi-class intertrochanteric femur fracture detection from CT images based on AO/OTA classification using faster R-CNN-BO method". In: ISSN: 1214-0287. DOI: 10.32725/jab.2020.013. URL: http://jab.zsf.jcu.czhttp//doi.org/10.32725/jab.2020.013.

Yu, Wei, Li, Maoqing, and Li, Xin (2012). "Fragmented skull modeling using heat kernels". In: *Graphical Models* 74.4, pp. 140–151.

Al-Zadjali, Najiba (2017). "Computer-aided diagnosis of complications of total hip replacement X-ray images". PhD thesis. Loughborough University.

Zhang, Kang, Yu, Wuyi, Manhein, Mary, Waggenspack, Warren, and Li, Xin (2015a). *3D Fragment Reassembly Using Integrated Template Guidance and Fracture-Region Matching.*

– (2015b). "3d fragment reassembly using integrated template guidance and fracture-region matching". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2138–2146.

Zhang, Zhengxin, Liu, Qingjie, and Wang, Yunhong (2018). "Road extraction by deep residual u-net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5, pp. 749–753.

Zhang, Zheyuan, Yao, Lanhong, Keles, Elif, Velichko, Yury, and Bagci, Ulas (2023). "Deep Learning Algorithms for Pancreas Segmentation from Radiology Scans: A Review". In: *Advances in Clinical Radiology.*

Zhou, Beibei, Willis, Andrew, Sui, Yunfeng, Anderson, Donald, Thomas, Thaddeus, and Brown, Thomas (2009). "Improving inter-fragmentary alignment for virtual 3D reconstruction of highly fragmented bone fractures". In: *SPIE medical imaging.* International Society for Optics and Photonics, pp. 725934–725934.

Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, Oliva, Aude, and Torralba, Antonio (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhou, Zongwei, Rahman Siddiquee, Md Mahfuzur, Tajbakhsh, Nima, and Liang, Jianming (2018). "Unet++: A nested u-net architecture for medical image segmentation". In: *Deep*

Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp. 3–11.

Zotti, Clement, Luo, Zhiming, Lalande, Alain, and Jodoin, Pierre-Marc (2018). "Convolutional neural network with shape prior applied to cardiac MRI segmentation". In: *IEEE journal of biomedical and health informatics* 23.3, pp. 1119–1128.