Automatic Quantification of Radiographic Knee Osteoarthritis Severity and Associated Diagnostic Features using Deep Convolutional Neural Networks

A. Joseph Antony, B.E., M.E.

A Dissertation submitted in fulfilment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

School of Electronic Engineering

Supervisors: Dr. Kevin McGuinness, Prof. Noel E O'Connor,

Dr. Kieran Moran

November 2017

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Sign: Student No.: 13211970

(A. Joseph Antony)

Date:

Acknowledgements

I express my heartfelt gratitude to my supervisor, Prof. Noel E O'Connor, for giving me an excellent opportunity to work under his supervision, for supporting my research directions and encouraging me throughout my PhD, and for being my true inspiration. I will be ever grateful to my supervisor Dr. Kevin McGuinness for his brilliant ideas, insightful feedbacks, sound advice, and continual support during my PhD. I would like to thank my supervisor, Dr. Kieran Moran for his excellent inputs, intriguing questions, and consistent encouragement.

I extend my sincere thanks to the Insight Centre for Data Analytics and Science Foundation Ireland for funding my research.

I thank all my friends and colleagues in DCU, for their companionship, for sharing the lighter moments and including me in all their celebrations.

I would like to gratefully acknowledge Prof. Sumam David, NITK, for motivating and supporting me to pursue my PhD. I am grateful to the Department of E & C and the management of NITK for providing me study leave for the period of my PhD.

On a personal note, I thank my parents (Antoni Guruz and Jayamary) and my sisters (Jancy and Gleeta) for being my pillars of strength and support.

I would like to extend my heartfelt thanks to everyone who helped and supported me throughout my PhD.

"I will give thanks to you, Lord, with all my heart; I will tell of all your wonderful deeds." (Psalm 9.1, NIV).

Contents

1 Introduction **1**

1.1 Chapter Overview 1

1.2 Knee Osteoarthritis 1

1.2.1 Diagnostic Features 2

1.2.2 Clinical Significance of Knee OA Studies 3

1.2.3 Radiographic Classification of Knee OA 3

1.3 Motivation 6

1.4 Hypotheses and Research Questions 8

1.5 Research Objectives 14

1.6 Thesis Outline 16

2 Related Work and Background **18**

2.1 Introduction 18

2.2 Detecting Knee Joints in Radiographs 19

2.2.1 Manual Methods 20

2.2.2 Semi-automatic Methods 21

2.2.3 Automatic Methods 22

2.3 Assessing Radiographic Knee OA Severity 23

2.3.1 Quantitative Analysis 24

2.3.2 Automatic Classification 25

2.4 Discussion 27

2.5 Public Knee OA Datasets 28

2.6 Deep Learning 30

2.6.1 Convolutional Neural Networks 32

2.6.2 Feature Learning 34

2.6.3 Transfer Learning and Fine Tuning 35

2.7 Chapter Summary 36

3 Baseline Methods **39**

3.1 Introduction 39

3.2 Dataset 41

3.3 Automatic Detection of Knee Joints 42

3.3.1 Template Matching 43

3.3.2 SVM-based Detection 46

3.4 Classifying Knee OA Images 49

3.4.1 WNDCHRM Classification 50

3.4.2 Classification using Hand-crafted Features 52

3.4.3 Classification using CNN Features 54

3.5 Chapter Summary 57

4 Automatic Localisation of Knee Joints **59**

4.1 Introduction 59

4.2 Localisation with Reference to Knee Joint Centre 60

4.2.1 Dataset and Ground Truth Generation 61

4.2.2 Training Fully Convolutional Neural Networks 61

4.2.3 Quantitative Evaluation 64

4.2.4 FCN with Pooling and Up-sampling Layers 66

4.2.5 Best Performing FCN for Detecting the Knee Joint Centres . 68

4.2.6 Error Analysis 70

4.2.7 Automatically Extracting the Knee Joints 72

4.2.8 Localisation Results 73

4.2.9 Limitations of this Method 74

4.3 Localising the Region of Interest 76

4.3.1 Dataset and Ground Truth 76

4.3.2 Training the FCN 78

4.3.3 Quantitative Evaluation 79

4.3.4 Qualitative Evaluation 80

4.3.5 Error Analysis 81

4.3.6 Extracting the Knee Joints 81

4.4 Conclusion 83

4.5 Chapter Summary 84

5 Automatic Quantification of Knee OA Severity 86

5.1 Introduction 86

5.2 Off-the-shelf CNNs 88

5.2.1 Dataset 89

5.2.2 Classification using Pre-trained CNN Features 89

5.2.3 Classification using Fine-tuned CNNs 92

5.2.4 Regression using Fine-tuned CNNs 94

5.2.5 Discussion 96

5.3 Training CNNs from scratch 97

5.3.1 Dataset and Preprocessing 98

5.3.2 Initial Configuration 98

5.3.3 Training Process and Initial Results 99

5.3.4 Tuning Hyper-parameters 100

5.3.5 Training Off-the-shelf CNNs from Scratch 103

5.3.6 Best Performing CNN for Classification 107

5.3.7 Classification Results 110

5.3.8 Comparison of Manual and Automatic Localisations 111

5.3.9 Training CNNs for Regression 112

5.3.10 Comparison of Classification and Regression Results 114

5.4 Multi-objective Convolutional Learning 116

5.4.1 Initial Configuration 117

5.4.2 Tuning Hyper-parameters 119
5.4.3 Best Performing Jointly Trained CNN 121
5.4.4 Jointly Trained CNN Results 122
5.4.5 Results Comparison 124
5.4.6 Error analysis 125
5.4.7 Discussion 127
5.5 Ordinal Regression 128
5.5.1 CNN Configuration for Ordinal Regression 128
5.5.2 CNN Training 129
5.5.3 Results 131
5.6 Results Comparison 133
5.7 Chapter Summary 134

6 Automatic Quantification of Knee OA Diagnostic Features 136

6.1 Introduction 136
6.2 Dataset and Ground Truth 138
6.3 Training CNNs to Quantify JSN 140
6.3.1 CNN Configuration and Training Process 140
6.3.2 Classification Results for Lateral JSN 141
6.3.3 Classification Results for Medial JSN 142
6.4 Training CNNs to Quantify Osteophytes 143
6.4.1 Quantifying Femoral Osteophytes 144
6.4.2 Classification Results for Lateral Femoral Osteophytes 144
6.4.3 Classification Results for Medial Femoral Osteophytes 145
6.4.4 Discussion 145
6.5 Jointly Training a CNN to quantify JSN, Osteophytes and KL grades 146
6.5.1 Initial Configuration 147
6.5.2 Results from Initial Configuration 148
6.5.3 Tuning the Hyper-parameters 150

6.5.4 Joint Training for Classification and Regression 151

6.5.5 Best Performing Jointly trained CNN for Classification . . . 153

6.5.6 Jointly Trained CNN Classification Results 155

6.6 Comparison of Individual and Joint Training Results 157

6.7 Comparison of Classification Results to the OAI Reliability Readings 158

6.7.1 Multi-class Classification 159

6.7.2 Binary Classification 160

6.7.3 Comparison of Binary and Multi-class Classifications 161

6.8 An Automatic Knee OA Diagnostic System 162

6.9 Chapter Summary 164

7 Conclusion 165

7.1 Thesis Overview 165

7.2 Research Questions and Proposed Solutions 169

7.3 Research Contributions 173

7.4 Future Work 174

7.5 Concluding Remarks 176

Bibliography 178

List of Publications

[1] *Peer-reviewed*: Joseph Antony, Kevin McGuinness, Kieran Moran, Noel E O'Connor. Automatic Detection of Knee Joints and Quantification of Knee Osteoarthritis Severity using Convolutional Neural Networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)* proceedings pp.376-390, Springer, 2017.

[2] *Peer-reviewed*: Joseph Antony, Kevin McGuinness, Noel E O'Connor, Kieran Moran. Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks. In *23rd International Conference on Pattern Recognition (ICPR)* proceedings pp.1195-2000, IEEE, 2016.

[3] *Peer-reviewed*: Joseph Antony, Kevin McGuinness, Neil Welch, Joe Coyle, Andy Franklyn-Miller, Noel E O'Connor, Kieran Moran. An Interactive Segmentation tool for Quantifying Fat in Lumbar Muscles using Axial Lumbar-Spine MRI. *IRBM*, Elsevier, 37(1), pp.11-22. 2015.

[4] *Peer-reviewed*: Joseph Antony, Kevin McGuinness, Neil Welch, Joe Coyle, Andy Franklyn-Miller, Noel E O'Connor, Kieran Moran. Fat Quantification in MRI-defined Lumbar Muscles. In *4th International Conference on Image Processing Theory, Tools and Applications (IPTA)* proceedings pp.1-6, IEEE, 2014.

[5] *Peer-reviewed*: Neil Welch, Kieran Moran, Joseph Antony, Chris Richter, Joe Coyle, Eanna Falvey, and Andrew Franklyn-Miller. The effects of a free-weight-based resistance training intervention on pain, squat bio-mechanics and MRI-defined lumbar fat infiltration and functional cross-sectional area in those with chronic low back, *BMJ open sport & exercise medicine 1*, no. 1, BMJ Specialist Journals, 2015.

[6] *Workshop*: Kevin McGuinness, Eva Mohedano, Amaia Salvador, Zhenxing, Mark Marsden, Peng Wang, Iveel Jargalsaikhan, Joseph Antony, Xavier Giro-i-Nieto, Shin'ichi Satoh, Noel O'Connor and Alan Smeaton. Insight DCU at TRECVID 2015, *NIST TRECVID Workshop*, NIST, 2015.

List of Poster Presentations

[1] Joseph Antony, Kevin McGuinness, Kieran Moran, Noel E O'Connor. Automatically Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks. *Insight Augmented Human Demonstrator Showcase*, Croke Park, Dublin, 2017.

[2] Joseph Antony, Kevin McGuinness, Kieran Moran, Noel E O'Connor. Assessing Knee Osteoarthritis Severity in a Continuous Scale using Convolutional Neural Networks. *Insight Student Conference*, DCU, Dublin, 2016.

[3] Joseph Antony, Kevin McGuinness, Kieran Moran, Noel E O'Connor. Radiographic Knee Osteoarthritis Classification using Convolutional Neural Networks. *Insight Student Conference*, NUI, Galway, 2015.

[4] Joseph Antony, Kevin McGuinness, Noel E O'Connor, Kieran Moran. Automatic Classification of Knee Osteoarthritis Images. *Insight Student Conference*, UCD, Dublin, 2014.

List of Abbreviations and Acronyms

ANN	Artificial neural network
BME	Bone marrow edema
CAD	Computer aided diagnosis
CI	Confidence interval
CNN	Convolutional neural network
CRF	Conditional random fields
CT	Computer axial tomography
DBN	Deep belief network
FCN	Fully convolutional network
GMM	Gaussian mixture models
GLCM	Gray level co-occurrence matrix
HOG	Histogram of oriented gradients
IKDC	International knee documentation committee
ILSVRC	ImageNet large scale visual recognition challenge
IoU	Intersection over union
JI	Jaccard index
JSN	Joint space narrowing
JSW	Joint space width
kNN	k - Nearest neighbour
KIDA	Knee images digital analysis
KL	Kellgren and Lawrence
KOACAD	Knee osteoarthritis computer aided diagnosis
LBP	Local binary patterns
LDA	Latent Dirichlet allocation
mJSW	Minimum joint space width
MLP	Multi layer perceptron
MOST	Multicenter osteoarthritis study

MRF	Markov random fields
MRI	Magnetic resonance imaging
NIH	National institutes of health
OA	Osteoarthritis
OAI	Osteoarthritis initiative
OARSI	Osteoarthritis research society international
PCA	Principal component analysis
RELU	Rectified linear unit
ROC	Receiver operating characteristic
ROI	Region of interest
SGD	Stochastic gradient descent
SVM	Support vector machine
SVR	Support vector regression
VGG	Visual geometry group
WNDCHARM	Weighted neighbour distances using a compound hierarchy of algorithms representing morphology
WOMAC	Western Ontario and McMaster Universities osteoarthritis index

List of Tables

3.1 Detection of knee joint centres using template matching method. . . 46

3.2 Classification metrics of the SVM for detection. 48

3.3 Comparison of template matching and the proposed SVM-based method. 49

3.4 Results of WNDCHRM Classification. 51

3.5 Classification results of the proposed methods using hand-crafted features. 52

3.6 Classification results of WNDCHRM and the proposed methods using hand-crafted features. 53

3.7 Classification results of WNDCHRM and the proposed methods. . . 56

3.8 Multi-class classification using WNDCHRM and CNN features. . . . 57

4.1 Initial FCN Configuration for detecting the knee joint centres. 62

4.2 FCN for detecting the knee joint centres. 64

4.3 FCN with pooling and up-sampling layers. 66

4.4 FCN with 3 Convolution-Pooling stages for detecting the knee joint centres. 67

4.5 Detection accuracy of FCN based on Jaccard Index. 68

4.6 Best performing FCN for detecting the knee joint centres. 69

4.7 Detection accuracy of the best performing FCN. 70

4.8 Comparison of methods used for localising the centre of the knee joints 74

4.9 Comparison of automatic detection based on the Jaccard Index (JI). 80

5.1 Classification accuracy (%) achieved by the WNDCHRM and pre-trained CNN features. 91

5.2 Classification accuracy (%) achieved with the features extracted from fine-tuned BVLC Net. 93
5.3 MSE for classification and regression. 95
5.4 Comparison of classification performance using classification (left) and regression (right) losses. 96
5.5 Initial CNN configuration. 99
5.6 Classification results of the initial CNN configuration. 100
5.7 CNN architecture (CNN-1) after tuning hyper-parameters. 101
5.8 Classification results after tuning hyper-parameters. 102
5.9 CNN architecture (CNN-2) after tuning hyper-parameters. 103
5.10 AlexNet architecture. 104
5.11 Classification results of the fully trained AlexNet. 105
5.12 VGG-M-128 architecture. 107
5.13 Classification results of the fully trained VGG-M-128. 107
5.14 Best performing CNN for classifying the knee images. 109
5.15 Classification results of the best performing fully trained CNN. . . . 109
5.16 Classification results of the proposed methods and the existing methods.110
5.17 Comparison of classification results obtained using manually annotated knee joints to automatically localised knee joints. 112
5.18 Comparison of manual and automatic localisation performance. . . . 112
5.19 Results of the initial network trained for regression after rounding the predicted continuous grades. 114
5.20 Best performing CNN for regression of the knee images. 115
5.21 Comparison of classification and regression results. 115
5.22 Comparison of the regression and classification performances. 117
5.23 Initial configuration to jointly train a CNN for classification and regression outputs. 118
5.24 Jointly trained network for classification and regression outputs. . . 121
5.25 Jointly trained network for classification and regression outputs. . . 122

5.26 Results of the best performing jointly trained CNN for classification and regression outputs. 124
5.27 Comparison of results from jointly trained CNN and individually trained CNNs for classification and regression results. 124
5.28 CNN architecture for ordinal regression 130
5.29 Comparison of classification metrics from regression and ordinal regression. 133
5.30 Comparison of classification, regression and ordinal regression results. 134

6.1 Grade-wise distribution of the knee joints as per KL grades, JSN, and osteophytes for the radiographs in the combined OAI-MOST dataset. 140
6.2 CNN for classifying knee images based on lateral JSN. 141
6.3 Results for classifying knee images based on lateral JSN. 142
6.4 Results for classifying knee images based on medial JSN. 143
6.5 Results for classifying knee images based on lateral femoral osteophytes.144
6.6 Results for classifying knee images based on medial femoral osteophytes.145
6.7 Initial configuration of the jointly trained network for classification of knee images based on KL grades, JSN, and osteophytes. 148
6.8 Classification results of jointly trained CNN to classify knee images based on KL grades, JSN, and osteophytes. 150
6.9 Results of jointly trained CNN to classify femur medial osteophytes. 151
6.10 Best performing jointly trained CNN for classification of knee images based on KL grades, JSN, and osteophytes. 152
6.11 Classification results of the best performing jointly trained CNN to classify knee images based on KL grades, JSN, and osteophytes. . . . 157
6.12 Comparing the results from individually trained CNNs to the jointly trained CNN for classifying knee images based on KL grades, JSN, and osteophytes. 158

6.13 Comparison of the multi-class classification results to the OAI radiologic reliability readings in ordinal scale. 160

6.14 Comparison of the binary classification results to the OAI radiologic reliability readings in binary scale. 161

6.15 Comparing the results of multi-class classification to binary classification for classifying knee images based on KL grades, JSN, and osteophytes. 162

List of Figures

1.1 A healthy knee and a knee joint affected with OA. 2

1.2 Stages of knee OA. 4

1.3 The Kellgren and Lawrence grading system to assess the severity of knee OA. 5

2.1 Samples of bilateral PA fixed flexion knee OA radiographs. 29

2.2 The OAI baseline data set distribution based on KL grades. 30

2.3 The MOST data set distribution based on KL grades. 31

2.4 Some examples of the learned features from a CNN. 37

3.1 The OAI baseline data (200 X-ray images) distribution as per KL grades. 41

3.2 Samples of X-ray images before (left) and after (right) histogram equalisation. 42

3.3 A knee OA X-ray image with the region of interest: the knee joints. 43

3.4 Pre-selected knee joint centres (20×20 pixels) extracted from knee joint images for template matching. 44

3.5 Detecting the knee joint centres and extracting the knee joints. . . . 48

3.6 The VGG-16 layers network architecture. 55

4.1 Automatic localisation of knee joints with reference to the centre of the knee joints. 61

4.2 (a) An input X-ray image and (b) The binary mask annotations for knee joint centres. 62

4.3 An instance of input, ground truth and output (predictions) of FCN. 63

4.4 An input image, ground truth, and outcome of the final FCN. 65

4.5 Prediction of the FCN with max pooling and up-sampling layers. . . 66

4.6 Predictions of the FCN with 3 Convolution-Pooling stages. 67

4.7 Training and validation losses of the FCN. 70

4.8 Error analysis: X-ray images, ground truth, FCN output - weak detections 71

4.9 Error analysis: X-ray images, ground truth, FCN output - detections with noise. 72

4.10 A knee X-ray image with the detected centres and the extracted left and right knees. 74

4.11 Anomalies in the automatic extraction of the ROI. 75

4.12 The actual ROI for the knee joints in Figure 4.11. 75

4.13 Variations in the aspect ratio of the extracted knee joints. 76

4.14 The actual ROI for the extracted knee joints in Figure 4.13. 76

4.15 Automatic localisation of the Region of Interest. 77

4.16 (a) An input X-ray image and (b) The binary mask annotations for the region of interest. 78

4.17 Training and validation losses of the FCN. 79

4.18 An input X-ray image, ground truth and output prediction of the FCN. 79

4.19 Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.98, right knee with JI=0.888. . 80

4.20 Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.879, right knee with JI=0.969. . 81

4.21 Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.768, right knee with JI=0.984. . 81

4.22 Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.83, right knee with JI=0.398. The implants in the right knee is the reason for this localisation error. . . 82

4.23 Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.473, right knee with JI=0.837. The implants in the left knee is the reason for this localisation error. . . . 82

4.24 Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.887, right knee with JI=0.356. The noise in the right knee causes this localisation error. 82

4.25 Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.681, right knee with JI=0.488. The localisation error in this image is due to the variation in the imaging protocol. 83

4.26 Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.768, right knee with JI=0.507. The variations in the local contrast and luminance affects the localisations. 83

5.1 Learning curves:training and validation losses (left), and validation accuracy (right) during fine-tuning. 92

5.2 Learning curves: training and validation losses, and accuracies of the fully trained CNN 102

5.3 Learning curves: training and validation losses, and accuracies of the fully trained AlexNet. 106

5.4 Learning curves: training and validation losses, and accuracies of the fully trained VGG-M-128 network. 108

5.5 Learning curves: training and validation losses, and accuracies of the best performing fully trained CNN. 110

5.6 Learning curves: training and validation losses for the best performing CNN for regression. 116

5.7 Initial configuration to jointly train a CNN for classification and regression outputs. 118

5.8 Learning curves for (a) classification and (b) regression in jointly trained CNN. 120

5.9 Learning curves for (a) classification and (b) regression in jointly trained CNN. 123

5.10 Confusion matrix for the multi-class classification using the jointly trained CNN. 125

5.11 ROC for the multi-class classification using the jointly trained CNN. 126

5.12 Mis-classifications: grade 1 joints predicted as grade 0, 2, and 3. . . 126

5.13 Misclassification: other grade knee joints predicted as grade 1. . . . 126

5.14 An instance of more severe misclassification: grade 0 and grade 3. . 127

5.15 The CNN configuration for ordinal regression. 129

5.16 Learning curves for (a) classification and (b) ordinal regression . . . 131

5.17 The CNN configuration for ordinal regression. 132

6.1 A knee radiograph showing healthy cartilage in the lateral compartment and joint space narrowing (JSN) due to cartilage loss in the medial compartment. 139

6.2 A few instances of visible femoral and tibial osteophytes in the lateral and medial regions. 139

6.3 Initial configuration of the jointly trained network for classification of knee images based on KL grades, JSN, and osteophytes. 149

6.4 Learning curves for KL grades classification in jointly trained CNN. 150

6.5 Learning curves for KL grades classification in jointly trained CNN. 153

6.6 Learning curves for classification of (a) lateral JSN and (b) medial JSN in jointly trained CNN. 154

6.7 Learning curves for classification of (a) femur lateral and (b) femur medial osteophytes severity in jointly trained CNN. 155

6.8 Learning curves for classification of (a) tibia lateral and (b) tibia medial osteophytes severity in jointly trained CNN. 156

6.9 Comparison of the results to the OAI radiologic reliability readings. 160

6.10 An end-to-end knee OA diagnostic pipeline. 163

Abstract

"Automatic Quantification of Radiographic Knee Osteoarthritis Severity and Associated Diagnostic Features using Deep Convolutional Neural Networks"

A. Joseph Antony

Due to the increasing prevalence of knee Osteoarthritis (OA), a debilitating knee-joint degradation, and total joint arthoplasty as a serious consequence, there is a need for effective clinical and scientific tools to assess knee OA in its early stages. This thesis investigates the use of machine learning algorithms and deep learning architectures, in particular convolutional neural networks (CNN), to quantify the severity and clinical radiographic features of knee OA. The goal is to offer novel and effective solutions to automatically assess the severity of knee OA achieving on par with human accuracy. Instead of conventional hand-crafted features, it is proposed in this thesis that automatically learning features in a supervised manner can be more effective for fine-grained knee OA image classification.

The main contributions of this thesis are as follows. First, the use of off-the-shelf CNNs are investigated for classifying knee OA images through transfer learning by fine-tuning the CNNs. Second, CNNs are trained from scratch to quantify the knee OA severity optimising a weighted ratio of two loss functions: categorical cross entropy and mean-squared error. Third, CNNs are jointly trained to quantify the clinical features of knee OA: joint space narrowing (JSN) and osteophytes along with the KL grades. This improves the overall quantification of knee OA severity producing simultaneous predictions of KL grades, JSN and osteophytes. Two public datasets are used to evaluate the approaches, the OAI and the MOST, with extremely promising results that outperform existing approaches. In summary, this thesis primarily contributes to the field of automated methods for localisation and quantification of radiographic knee OA.

Chapter 1

Introduction

1.1 Chapter Overview

This chapter provides a general introduction to this thesis, presents the motivation for this research, the hypotheses and research questions, and outlines the structure of this thesis. Section 1.2 introduces knee osteoarthritis (OA), the diagnostic features of knee OA, the clinical significance, and the clinical evaluation of knee OA using the Kellgren and Lawrence (KL) grading scheme. Section 1.3 discusses the specific motivations underpinning this research. Section 1.4 presents the hypotheses and research questions derived from the previously reported work on early detection and assessment of knee OA severity. Section 1.5 lists the research objectives. Section 1.6 outlines the organisation and structure of this thesis.

1.2 Knee Osteoarthritis

Knee OA is a debilitating joint disorder that mainly degrades the knee articular cartilage. In general, knee OA is characterised by joint pain, cartilage wear, and bony growths. Knee OA has a high-incidence among the elderly, obese, and those with a sedentary lifestyle. In its severe stages, it causes excruciating pain and often leads to total joint arthoplasty. Early diagnosis is crucial for clinical treatments and pathology [1, 2].

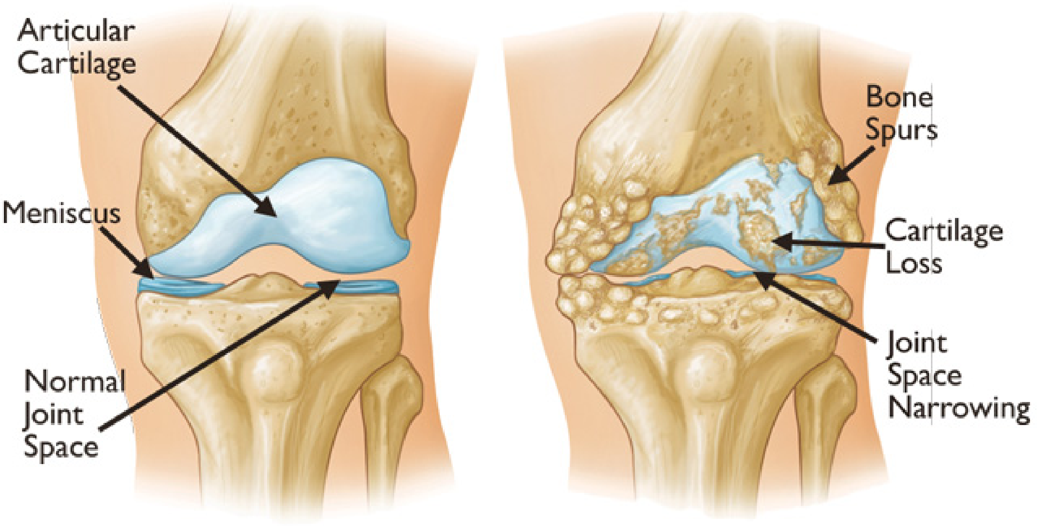


Figure 1.1: A healthy knee and a knee joint affected with OA.

Source: `http://orthoinfo.aaos.org/figures/A00389F02.jpg`

1.2.1 Diagnostic Features

Clinically, the major pathological features for knee OA include joint space narrowing, osteophytes formation, and sclerosis [1,3]. Figure 1.1 shows the anatomy of a healthy knee and a knee affected with osteoarthritis, and the characteristic features of knee OA: joint space narrowing (JSN) due to cartilage loss and bone spurs (osteophytes). The causes for knee OA include mechanical abnormalities such as degradation of articular cartilage, menisci, ligaments, synovial tissue, and sub-chondral bone.

The major clinical features; joint space narrowing and osteophyte formation, are easily visualised using radiographs [1, 4, 5]. Despite the introduction of several imaging methods such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound for augmented OA diagnosis, radiographs have traditionally been preferred [5,6], and remain as the main accessible tool and "gold standard" for preliminary knee OA diagnosis [2, 7]. However, many argue that MRI is the most useful imaging modality to study the structural variations and to visualise soft tissues, and that MRI provides the structural evidence of knee

OA [8]. Inspired by the previous successful approaches in the literature for early identification [2] and automatic assessment of knee OA severity [1,6,9], the focus is on radiographs in this thesis. More importantly, there are public datasets available that contain radiographs with associated ground truth. Public datasets for knee OA study, such as the OAI and the MOST datasets, provide radiographs with Kellgren and Lawrence (KL) scores, and the OARSI[1] readings for distinct knee OA features such as JSN, osteophytes, and sclerosis. Section 1.2.3 discuses the KL scores in detail. The details of OARSI readings are discussed in Chapter 6.

1.2.2 Clinical Significance of Knee OA Studies

Due to the increasing prevalence of knee OA, diminishing health-related quality of life, and total joint arthoplasty as a serious consequence, there is a growing need for effective clinical and scientific tools for early detection of knee OA reliably [1,2,7]. Early identification of knee OA and assessment of the severity are crucial for pathology, clinical decision making, and to study disease progression [6].

As per a recent study [10], more than 250 million people across the globe are affected by knee OA alone. A study [11] on future projections of total hip and knee arthoplasty in the UK estimates the total primary hip and knee replacement counts in 2035 at 439,097 and 1,219,362 respectively. The National Institutes of Health (NIH) has sponsored a research project called the Osteoarthritis Initiative (OAI) to develop a public domain resource to facilitate knee OA research, to identify and validate knee OA biomarkers that will help to better understand how to prevent and treat knee OA.

1.2.3 Radiographic Classification of Knee OA

Knee OA develops gradually over years and progresses in stages. In general, the severity of knee OA is divided into five stages. The first stage (stage 0) corresponds to normal healthy knee and the final stage (stage 4) corresponds to

[1] Osteoarthritis Research Society International

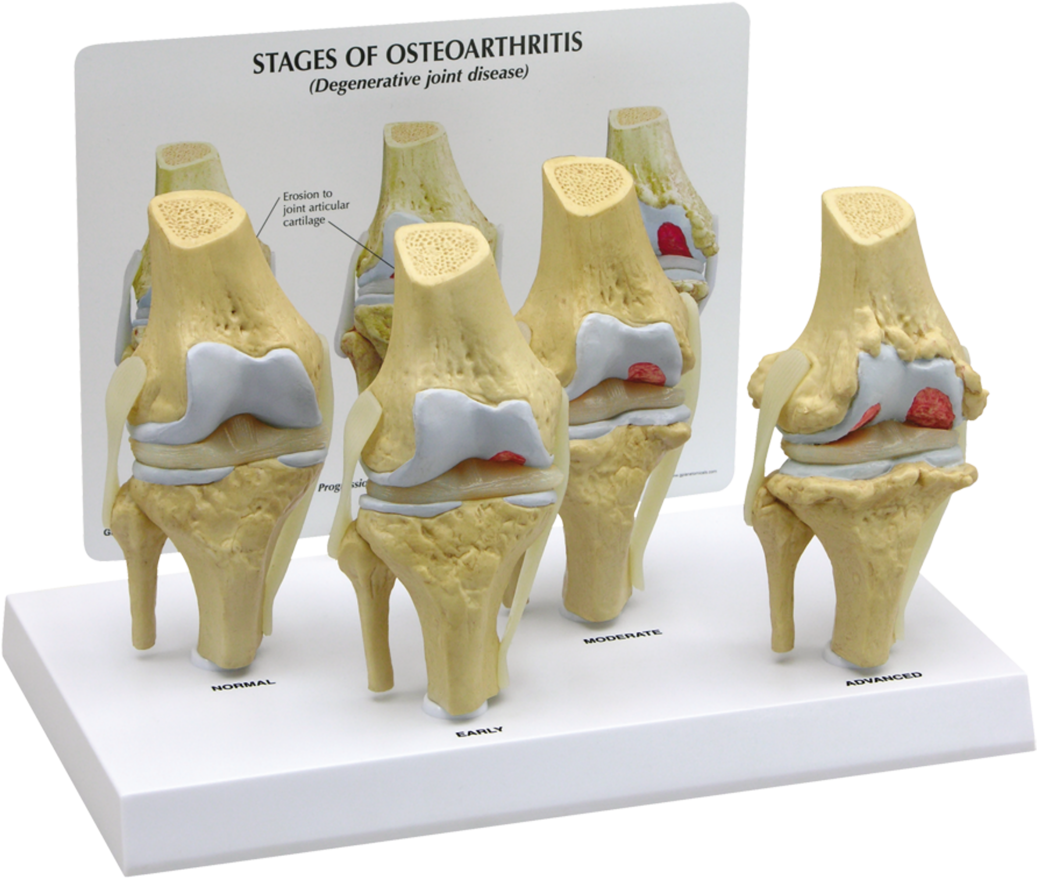

STAGES OF OSTEOARTHRITIS
(Degenerative joint disease)
Erosion to joint articular cartilage
NORMAL
EARLY
MODERATE
ADVANCED

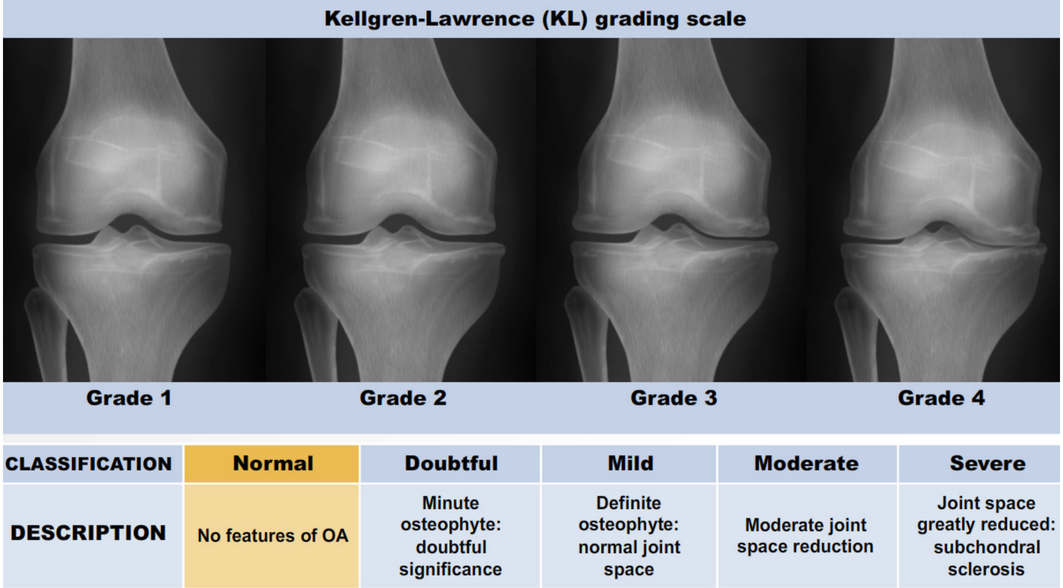
Figure 1.2: Stages of knee OA.

Source: `https://www.anatomynow.com/products/human_osteoarthritic_knee_model`

the most severe condition. Figure 1.2 illustrates the different stages of knee OA severity. The most commonly used systems for grading knee OA are the International Knee Documentation Committee (IKDC) system, the Ahlback system, and the Kellgren & Lawrence system[2]. The other widely used non-radiographic knee OA assessment system is WOMAC[3], which measures pain, stiffness, and functional limitation. As the public datasets used in the studies in this thesis are provided with KL grades, they are used as the ground truth to classify the knee OA X-ray images.

[2] `https://radiopaedia.org/articles/kellgren-and-lawrence-system-for/-classification-of-osteoarthritis-of-knee/`

[3] `http://www.womac.org/womac/index.htm`



CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Figure 1.3: The Kellgren and Lawrence grading system to assess the severity of knee OA.

Source: `http://www.adamondemand.com/clinical-management-of-osteoarthritis/`

Kellgren and Lawrence Scores

The Kellgren and Lawrence (KL) grading scale was approved by the World Health Organisation (WHO) as the reference standard for cross-sectional and longitudinal epidemiologic studies [4, 5, 12, 13]. The KL grading system is still considered the gold standard for initial assessment of knee osteoarthritis severity in radiographs [1, 4, 7, 14]. Figure 1.3 shows the KL grading system. The KL grading system categorises knee OA severity into five grades (grade 0 to 4). The KL grading scheme for quantifying knee OA severity from X-ray images is defined as follows.

- Grade 0: absence of radiographic features (cartilage loss or osteophytes) of OA.
- Grade 1: doubtful joint space narrowing (JSN), osteophytes sprouting, bone marrow oedema (BME), and sub-chondral cyst.
- Grade 2: visible osteophytes formation and reduction in joint space width on the antero-posterior weight-bearing radiograph with BME and sub-chondral

cyst.

- Grade 3: multiple osteophytes, definite JSN, sclerosis, possible bone deformity.
- Grade 4: large osteophytes, marked JSN, severe sclerosis, and definite bone deformity.

1.3 Motivation

Currently, experienced clinicians assess knee OA severity by grading the knee joints in X-ray images [15]. The most commonly used gradings like the KL grading scheme, and Ahlback system, use distinctive grades (0 to 4). However, clinical features of knee OA are continuous in nature, and attributing distinctive grades is the subjective opinion of the graders. There are also uncertainties and variations in the subjective gradings. There is a need for automated methods to overcome the limitations arising from this subjectivity, and to improve the reliability in the measurements and classifications [15].

The automatic assessment of knee OA severity has been previously approached in the literature as an image classification problem [2, 7, 9], with the KL grading scale as the ground truth. WNDCHARM, a multi purpose biomedical image classifier was used to classify knee OA images [9, 14]. High binary classification accuracies (80% to 91%) have been reported using the WNDCHARM classifier for classifying the extreme stages: grade 0 (normal) vs grade 4 (severe), grade 0 vs grade 3 (moderate). However, the classification accuracies of the images belonging to successive grades are low (55% to 65%) and the multi-class classification accuracy is low (35%). The overall classification accuracies of knee OA needs improvement for real-world computer aided diagnosis [1, 2, 7].

Radiographic features detected and learned through a computer-aided analysis can be useful to quantify knee OA severity and to predict the future development of knee OA [2]. Instead of manually designing features, the author proposes that learning feature representations using deep learning architectures can be a more effective approach for the classification of knee OA images.

Traditionally, hand-crafted features based on pixel statistics, object and edge statistics, texture, histograms, and transforms, are typically used for multi purpose medical image classification [4, 9, 14]. However, these features are not efficient for fine-grained classification such as classifying successive grades of knee OA images. Manually designed or hand-engineered features often simplify machine learning tasks. Nevertheless, they have a few disadvantages. The process of engineering features requires domain related expert knowledge and is often very time consuming [16]. These features are often low-level as prior knowledge is hand-encoded and features in one domain do not always generalise to other domains [17]. The next logical step is to automatically learn effective features for the desired task.

In recent years, learning feature representations is preferred to hand-crafted features, particularly for fine-grained classification, because rich appearance and shape features are essential for describing subtle differences between categories [18]. Feature learning correspond to techniques that learn to transform raw data input to an effective representation for further higher-level processing such as classification, automatic detection, and segmentation. Feature learning approaches provide a natural way to capture cues by using a large number of code words (sparse coding) or neurons (deep networks), while traditional computer vision features, designed for basic-level category recognition, may eliminate many useful cues during feature extraction [18]. Deep learning architectures are multi-layered and they are used to learn feature representations in the hidden layer(s). These representations are subsequently used for classification or regression at the output layer. Feature learning is an integral part of deep learning [16, 19].

Even though many deep learning architectures have been proposed and have existed for decades, in recent times convolutional neural networks (CNN) have become highly successful in the field of computer vision [20, 21]. AlexNet [22] won the ILSVRC[4] in 2012 by a large margin. CNNs have since become more popular, widely-used and highly-successful in computer vision tasks such as image recognition, automatic detection and segmentation, content based image retrieval, and video classification [20]. Apart from computer vision tasks, CNNs are finding applications in natural language processing, hyper-spectral image processing, and medical image analysis [20, 23]. Recently, CNNs have become successful in medical applications such as knee cartilage segmentation in MRI scans [24], brain tumour segmentation in magnetic resonance imaging (MRI) scans [25], multi-modality iso-intense infant brain image segmentation [26], pancreas segmentation in CT images [27], and neuronal membrane segmentation in electron microscopy images [28]. Inspired by these success stories, the author proposes CNNs for classification of knee OA images and to improve the quantification of knee OA severity and knee OA diagnostic features. The author believes that this can lead to build a real-world knee OA diagnostic system that outperforms the existing approaches.

1.4 Hypotheses and Research Questions

Based on the previously reported work in the early detection and computer aided diagnosis of knee OA severity, and the preliminary investigations, the author suggests the following hypotheses.

H1. *Learning feature representations and classification using supervised deep learning is more effective for assessing the severity of knee OA than conventional classification using hand-crafted features.*

H2. *Evaluating the automatic knee OA predictions using a continuous distance-based metric like mean squared error instead of classification accuracy is more appropriate and KL grades predictions can be approached as a regression problem. Training a CNN for optimising a*

[4] ImageNet Large Scale Visual Recognition Challenge

weighted ratio of two loss functions for simultaneous classification and regression can improve the accuracy of quantifying knee OA severity.

H3. *Jointly training a CNN for quantifying the clinical diagnostic features of knee OA such as joint space narrowing (JSN) and osteophytes, along with the KL grades will improve the overall quantification of knee OA severity.*

H4. *Using these improvements it is possible to build a computer aided diagnostic system to assess knee OA that is sufficiently accurate for practical application.*

Research Questions

From the hypotheses, the author outlines the following research questions and discusses how these questions are addressed by identifying potential solutions.

RQ1. What is the most efficient method for localising the ROI; i.e. the knee joint regions in X-ray images, in terms of speed and accuracy that also supports feature learning and classification using CNNs?

Radiologists and medical practitioners examine knee joint regions only in X-ray images for assessing knee OA severity and according to the literature, the region of interest (ROI) for classification is only the knee joint regions, i.e. the left and right knees joints [1, 4, 7, 9, 14]. Hence, detecting and localising the knee joint regions is an essential pre-processing step before classification. Automatic methods are preferable for large datasets such as the OAI and the MOST. The following methods are investigated to automatically localise the knee joints in X-ray images:

- Shamir et al. [2] proposed the template matching approach for automatic detection of knee joints. First, this method is implemented as a baseline in this thesis. (Section 3.3.1)

- Next, a SVM-based approach using Sobel horizontal image gradients as features to automatically localise the knee joints is investigated. (Section 3.3.2)
- A novel deep learning based approach is proposed for localisation and a fully convolutional network is trained. (Chapter 4)

RQ2. Instead of using hand-crafted features, is it possible to learn effective feature representations using a supervised deep learning method, in particular a convolutional neural network (CNN), for efficient and accurate fine-grained classification of knee OA images?

Previous approaches for early diagnosis and assessment of knee OA severity have used several hand-crafted features [1, 2, 4, 9, 29, 30] and conventional classification techniques such as SVM [30], k-nearest neighbour classifier [29], weighted nearest neighbour classifier [2, 9], random forest classifiers [15], and even artificial neural networks (ANN) [31, 32]. As a baseline, the state-of-the-art hand-crafted features successful in other computer vision tasks for classification of knee OA images are investigated. Next, supervised feature learning using CNNs for efficient classification of knee OA images is investigated, instead of using hand-crafted features. Deep learning based methods require large training data to generalise well after training the networks. Existing CNNs are trained with datasets like ImageNet [33] that contains more than 1 million images. Using deep learning based methods for computer aided diagnostics involves two challenges: lack of training data in comparison to datasets like ImageNet and it is unclear if deep learning methods will work well due to domain shift. The following investigations and solutions are proposed to address this research question:

- As a baseline approach, WNDCHARM proposed by Shamir et al. [9,34] to classify the knee images is implemented. This will be used to benchmark the classification results. (Section 3.4.1)

- The features extracted from off-the-shelf CNNs are investigated to classify knee OA images. (Section 3.4.3)
- The off-the-shelf pre-trained CNNs are fine-tuned through transfer learning for knee OA images classification. (Section 5.2)
- CNNs are trained from scratch for classifying knee OA images. (Section 5.3)

RQ3. As knee OA is progressive, can the categorisation of knee OA images be approached as a regression problem instead of classification?

Existing work on automatic assessment of knee OA severity treats it as an image classification problem, assigning each KL grade to a distinct category [2]. To date, evaluation of automatic KL grading algorithms has been based on binary and multi-class classification accuracy with respect to these discrete KL grades [1,7,14]. KL grades are not, however, categorical, but rather represent an ordinal scale of increasing severity. The quantisation of the KL grades to discrete integer levels is essentially an artefact of convenience; the true progression of the disease in nature is continuous, not discrete. The author justifies the use of regression of knee OA assessment and proposes the following methods to address this research question:

- The author argues it is more appropriate and useful to assess the accuracy of automatic knee OA predictions using a continuous distance-based metric like mean squared error than it is to use classification accuracy. The pre-trained CNNs are fine-tuned using both classification loss and regression loss. (Section 5.2)
- It is shown that a CNN fine-tuned with regression loss improves the classification accuracy in comparison to the CNN fine-tuned with classification loss. (Section 5.2.5)

RQ4. Can a CNN trained with a weighted ratio of two loss functions such

as categorical cross entropy and mean squared error improve the assessment of knee OA severity?

As pointed out before, it is more appropriate to use mean squared error as an evaluation metric instead of classification accuracy. This leads to the formulation of the prediction of KL grades as a regression problem. Furthermore, to obtain a better learning representation the networks are trained to learn using a weighted ratio of two loss functions: categorical cross entropy for classification and mean-squared error for regression. This question is solved with the following experiments:

- A CNN is jointly trained for simultaneous classification and regression of knee OA images. (Section 5.4)
- It is shown that there is an improvement in the classification performance of this jointly trained CNN in comparison to the CNN only trained for classification. (Section 5.4.5)

RQ5. Can ordinal regression be applied to automatically assess knee OA severity? How does this improve the overall assessment of knee OA severity?

Ordinal regression is investigated as a next step to further improve the quantification of knee OA severity. Ordinal regression is useful in particular to classify patterns using a categorical scale which shows a natural order between the labels [35, 36]. Ordinal regression[5] can be considered as an intermediate problem between classification and regression. The author believes that the KL grades prediction using ordinal regression can further improve the classification performance by reducing the margin of error (mean squared error) considering the progressive nature of knee OA and the ground truth or labels for training a CNN i.e. the KL grades are in ordinal scales

[5] *https://en.wikipedia.org/wiki/Ordinal_regression*

(0–4). The following is investigated and experimented to address this research question:

- An ordinal regression configuration is introduced for classifying knee OA images. (Section 5.5.1)
- A CNN is trained from scratch for ordinal regression of knee OA images (Section 5.5.2)
- It is shown that the CNN for ordinal regression gives better performance in comparison to the normal regression of knee OA images. (Section 5.5.3)

RQ6. Can jointly training a CNN for quantifying knee OA clinical features such as JSN and osteophytes along with KL grades further improve the overall quantification of knee OA severity?

Previous studies on early diagnosis and automatic assessment of knee OA severity [1, 2, 4, 7, 9, 14] have mainly focused on classifying X-ray images using KL grades as the ground truth. However, some studies claim that there have been differences in descriptions of KL grades [5, 12, 13, 37, 38]. Furthermore, there are studies claiming that KL grades are neither highly accurate nor reliable for radiographic classification of knee OA [39–42]. After a detailed study on the various grading scales used for knee OA assessment, Sheehy et al. [39] and Shamir et al. [42] claim that OARSI grading is more accurate and highly reliable for individual OA feature assessments. Thus, the prediction of clinical features of knee OA based on the OARSI grading system is investigated in this thesis. The investigations and experiments to address this research question are as follows.

- CNNs are trained from scratch to quantify lateral and medial JSN individually. (Section 6.3, Chapter 6)

- CNNs are trained separately to quantify femoral and tibial osteophytes in lateral and medial compartments. (Section 6.4)
- CNNs are jointly trained for quantifying KL grades, JSN and osteophytes to explore further improvement in the overall quantification of knee OA. (Section 6.5)

RQ7. How well do the results agree with the gold standard for assessing knee OA? Can the proposed methods be applied in practical computer aided diagnosis (CAD) of knee OA?

The suitability of the proposed methods are investigated for a real-world knee OA CAD system by comparing the performance of the proposed methods to the existing methods and the gold standard. The Cohen's kappa statistics for inter-rater agreement is used for this.

- Cohen weighted kappa values are calculated for the classification results with 95% confidence intervals (CI) to find the inter-rater agreement between the CNN predictions and the ground truth (KL grades). (Section 6.7)
- The performance of the proposed system is compared to the OAI kXR SQ reliability reading (BU), which is considered the gold standard for knee OA assessment. The weighted kappa values are used for this. (Section 6.7)
- An end-to-end pipeline combining the FCN for localising the knee joints and the CNN jointly trained for quantifying knee OA severity is developed. (Section 6.8)

1.5 Research Objectives

First, the objective is to present an up-to-date review of the literature and the state-of-the-art in the early detection and computer aided diagnosis of knee osteoarthritis. Next, by experimentally investigating and attaining suitable solutions to the research questions, the hypotheses will ideally be shown to hold true and the following research objectives will be achieved.

- Developing an efficient technique to automatically detect and localise the ROI: the knee joint regions in the X-ray images.
- Developing and evaluating a supervised deep learning framework using a CNN to classify the localised knee joint regions in an ordinal scale based on KL grades.
- Testing a continuous distance-based metric like mean squared error instead of classification accuracy to assess the automatic quantification of knee OA severity to examine if this improves the knee OA assessment.
- Training a CNN to quantify knee OA severity using regression in a continuous scale instead of a nominal scale, as knee OA is progressive by nature.
- Investigating joint training of a CNN with a weighted ratio of two loss functions such as categorical cross entropy and mean squared error for simultaneous multi-class classification and regression outputs.
- Investigating ordinal regression to improve the quantification of knee OA severity in a continuous scale.
- Jointly training a CNN following the multi-objective convolutional learning approach to quantify JSN, and osteophytes along with KL grades. The objective is to improve the overall quantification results of knee OA severity based on KL grades, JSN, and osteophytes.
- Comparing the performance of the proposed methods to the existing methods and the gold standard in knee OA assessment, and identifying the practical implications for the real-world use of the proposed methods.

1.6 Thesis Outline

The chapters in this thesis are organised as follows.

Chapter 2 provides an overview of the background, a comprehensive summary of the related work, and a critical analysis of the state-of-the-art in computer aided diagnosis of knee OA. The first part of the chapter reviews the existing manual and automatic methods for detecting the knee joints in radiographs, and emphasises the limitations of these methods. The next part discusses the various methods used for assessing knee OA severity, and describes in more detail WNDCHARM, the multi purpose medical image classifier. The chapter concludes with a comprehensive review of machine learning algorithms and deep learning architectures, in particular convolutional neural networks that are useful for quantifying knee OA severity.

Chapter 3 presents the preliminary work and the baseline methods used for automatic detection of knee joints in the radiographs and the classification of the localised knee joints. The first part reports on two methods: template matching and a SVM-based method to automatically detect and extract the knee joints. The next part presents the WNDCHARM implementation, and the proposed methods to classify knee OA images using hand-crafted features and CNN features.

Chapter 4 sets forth the proposed approaches for automatic detection of knee joints using deep learning. The first part presents an approach for automatically detecting the centre of knee joints using a fully convolutional network (FCN) and extracting the knee joint regions with reference to the detected centres. The next part describes a method to automatically detect the region of interest (ROI): knee joint regions directly, and also discusses the advantage of this method over the previous method.

Chapter 5 details the deep learning approaches for automatically quantifying knee OA severity. Three approaches are presented for this 1) fine tuning off-the-shelf pre-trained CNNs through transfer learning, 2) training a CNN from scratch for classifying knee OA images, 3) jointly training a CNN for classification and regression, and 4) training a CNN for ordinal regression. This chapter is concluded with a comparative analysis of all the results and a discussion.

Chapter 6 presents the automatic quantification of knee OA diagnostic features: joint space narrowing (JSN) and osteophytes. Following this, the joint training of a CNN for quantifying knee OA severity and the clinical features are discussed. This chapter concludes with the proposal to develop a practical knee OA diagnostic system and the practical implications to build this system.

Chapter 7 concludes this thesis by analysing the current work and summarising the research methodology, discussing the solutions to the research questions, highlighting the contributions, and providing future directions of research based on the proposed methods.

Chapter 2

Related Work and Background

2.1 Introduction

This chapter presents a comprehensive review of the literature and the state-of-the-art in assessing knee OA severity. It introduces the necessary background; i.e. learning feature representations, and quantifying knee OA severity using deep convolutional neural networks (CNN).

The automatic assessment of knee OA severity from radiographs has been approached as an image classification problem [2, 7, 9]. According to the literature, in the general approach to automatically assess knee OA severity, the first step is to localise the region of interest (ROI) that is to detect and extract the knee joint regions from the radiographs, and the next step is to classify the localised knee joints. First, the different approaches for detecting (or localising) the knee joint regions in the radiographs are outlined. Next, the approaches in the literature to assess knee OA severity are investigated and the focus is on the automated methods. Following this, the key pathological features of knee OA are introduced and the state-of-the-art methods for quantifying radiographic knee OA clinical features are reviewed. Also, the background and the state-of-the-art deep learning methods and architectures used in this thesis are outlined. This chapter concludes with a discussion outlining the limitations in the state-of-the-art methods on automatic detection of knee joints and automatic assessment of knee OA severity, and how these limitations can be addressed.

2.2 Detecting Knee Joints in Radiographs

Despite the introduction of several imaging modalities such as MRI, CT, and ultrasound for augmented OA diagnosis, radiography (X-ray) has been traditionally preferred, and remains the main accessible tool and "gold standard" for preliminary knee OA diagnosis [1, 7, 9]. The main pathological features of knee OA such as loss of joint cartilage, reduction in joint space width, and osteophytes (bone spurs), can be easily visualised and examined in plain radiographs [5, 6]. Advanced imaging modalities such as MRI and CT may be required when the clinical investigations from radiographs are inconclusive and do not give clear reasons for joint pain [6]. Expert radiologists specifically examine the knee joint regions in radiographs for joint space narrowing and osteophytes, for knee OA diagnosis [2].

There are several approaches in the literature for detecting and segmenting knee joints and specific parts of knee such as cartilage, menisci, and bones structures from 3D MRI and CT scan images [43, 44]. Nevertheless, the existing approaches are less accurate for automatically detecting the knee joints in radiographs [43, 45]. According to the literature, detecting knee joints remains a challenging task [43, 46]. Before the review of specific methods for knee joint segmentation, an overview of the segmentation approaches used in the medical field is presented.

In general, the medical image segmentation approaches are categorised into two basic groups: pixel-based and geometry-based [43,45]. The pixel-based segmentation methods include thresholding, region-growing, and region-merging. The geometry-based segmentations include active shape models, active appearance models, and deformable models such as snakes and active contours [43, 45]. There are other segmentation approaches that are classified as hybrid methods like atlas-guided approaches, clustering approaches, Markov random field models, and artificial neural networks (ANN) based approaches [43]. These hybrid methods need to be integrated with other segmentations to build a complete segmentation framework [43]. Some of these methods are used for detecting knee joints in radiographs.

In this thesis, automated methods for detecting knee joints in radiographs are investigated. For this reason, the author limits the scope of review to the approaches based on radiographs. The advantages of automatic methods are discussed and the need to investigate such methods are emphasised.

Previous approaches in the literature that investigate the knee joints in radiographs can be categorised into manual, semi-automatic, and fully automatic, based on the level of manual intervention required [43, 44]. The following sections review each of these approaches in detail.

2.2.1 Manual Methods

Expert radiologists or trained physicians visually examine the knee joint regions and trace the structures using simple image processing and computer vision-based tools in radiographs, and may even use CAD-based measurements for assessing knee OA severity [43]. The expert knowledge-based manual segmentations are useful to build an atlas or template of anatomical structures, which are used to develop advanced interactive and automatic segmentation methods [43]. The knee joints labelled manually are reliable and are often used as ground truth for evaluating automatic methods [47, 48]. Nevertheless, such manual methods are subjective, highly experience-based, and they are laborious and time-consuming when a large number of subjects are to be examined.

There are previous studies in the literature that use manually-defined ROIs (knee joints) in radiographs for assessing knee OA severity. Hirvasniemi et al. quantified the differences in bone density using texture analysis and local binary patterns (LBP) in plain radiographs to assess knee osteoarthritis [49]. Woloszynski et al. developed a signature dissimilarity measure for the classification of trabecular bone texture in knee radiographs [50]. In both these methods, the ROIs are manually marked and ROIs are extracted for texture analysis.

2.2.2 Semi-automatic Methods

Semi-automatic or interactive methods are developed to minimise manual interventions by automating essential steps in the detection and segmentation process [44, 51]. These methods often include manual initialisations with low-level image processing, followed by manual evaluations and corrections of the results [52]. The main advantage of the semi-automatic methods are flexibility in manual intervention that allow incorporating expert knowledge plus the use of advanced computer vision-based tools to automate the essential steps. An expert may improve the detection and segmentation performance through tuning the essential parameters for instance seed region and threshold values in region growing, initial shape of active models, delineating the required contour [43] to define the region of interest. However, these methods may not be reproducible due to inter-observer or inter-user variations and there is a possibility of oversight or human error in the manual evaluations.

There are some knee OA studies in the literature which use semi-automatic methods to detect the knee joints in radiographs. Knee OA computer aided diagnosis (KOACAD) proposed by Oka et al. [1] is an interactive method to measure the joint space narrowing, osteophytes formation and joint angulation in radiographs. In KOACAD, a Roberts filter is used to obtain the rough contour of tibia and femur bone structures and a vertical neighbourhood difference filter is used to identify points with high absolute values of difference of scales. The centre of all the points is calculated and a rectangular region around the centre, of size 480 × 200 pixels, is selected as the knee joint region. This system has provided accurate assessment of structural severity of knee OA after detecting the knee joint regions. However, human intervention is required for plotting various lines for the measurement and automatic detection is not feasible with this system.

Knee images digital analysis (KIDA) is a tool to analyse knee radiographs interactively, proposed by Marijnissen et al. [3]. KIDA quantifies the individual radiographic features of knee OA like medial and lateral joint space width (JSW) measurements, subchondral bone densities and osteophytes. This interactive tool can only be used by experts for quantitative measurements and requires expert intervention for objective quantitative evaluation.

Duryea et al. [53] proposed a trainable-rule based algorithm (software) to measure the joint space width between the edges of femoral condyle and the tibial plateau on knee radiographs. Contours marking the edges of femur and tibia are automatically generated. This interactive method can be used to monitor joint space narrowing and the progression of knee osteoarthritis.

2.2.3 Automatic Methods

Automatic segmentation methods have become an essential part of computer aided diagnosis and clinical decision support systems [46]. These methods are fast and accurate, and they are highly beneficial in clinical trials and pathology [43]. According to the literature, there have been multiple attempts to automatically localise knee joints in radiographs. Nevertheless, this task still remains a challenge.

Podsiadlo et al. [54] proposed an automated system for the prediction and early diagnosis of knee OA. In this approach, active shape models and morphological operations are used to delineate the cortical bone plates and locate the ROIs in radiographs. This approach is developed for selection of tibial trabecular bone regions in the knee joints as ROIs. Nevertheless, this approach can be extended to localise the entire knee joint. A set of 40 X-ray images are used for training and 132 X-ray images are used for testing in this method. The automatic detections in this method are compared to the gold standard, which contains manually annotated ROIs from the expert radiologists and the similarity indices (SI) are calculated. This method achieves SI of 0.83 for the medial and 0.81 for the lateral regions of the knee joints.

Shamir et al. [7] proposed template matching for automatic knee joint detection in radiographs. Template matching uses predefined joint centre images as templates and calculates Euclidean distances over every patch in an X-ray image using a sliding window. The image patch with the shortest distance is recorded as the detected knee joint centre. After detecting the centre, an image segment of 700×500 pixels around the centre is extracted as the knee joint region. The X-ray images from BLSA dataset are used in this method. In total 55 X-ray images from each grade are used for the experiments, such that 20 images from each grade for training and 35 images from each grade for testing. Shamir et al. reported that template matching was successful in finding the knee joint centres in all the X-ray images in their dataset.

Anifah et al. [55] investigated template matching and contrast-limited adaptive histogram equalisation for detecting knee joints and quantifying joint space area. In total 98 X-ray images are used in this method. The detection accuracy achieved by this method varies from 83.3% to 100% for the left knees and 60.4% to 100% for the right knees. Template matching is a simple and relatively fast method. However, this method is ad hoc, entirely based on the set of templates used and is unlikely to generalise well for larger datasets.

Recently, Tuilpin et al. [46] investigated a SVM-based method to automatically localise knee joints in plain radiographs. This method uses knee anatomy-based region proposals, and the best candidate region from the proposals are selected using histogram of oriented Gradients (HOG) as feature descriptors and a SVM. This method generalises well in comparison to the previous methods and shows reasonable improvement in automatic detections with mean intersection over union (IOU) 0.84, 0.79 and 0.78 on the public datasets MOST, Jyvaskyla, and OKOA.

2.3 Assessing Radiographic Knee OA Severity

The key pathological features of knee OA include joint space narrowing (JSN), osteophytes (bone spurs) formation, and sclerosis (bone hardening) [1, 3]. All these features are implicitly integrated in composite scoring systems like Kellgren & Lawrence (KL) grading system to quantify knee OA severity [1, 3] and the OARSI readings provide the gradings of distinct knee OA features. There are two common approaches in the literature for assessing knee OA severity in plain radiographs: 1) quantifying the distinct pathological features of knee OA and 2) automatic classification based on composite scoring systems such KL grades.

2.3.1 Quantitative Analysis

The most conventional system to assess radiographic knee OA severity has been KL gradings [1, 2, 14]. Nevertheless, some researchers [1, 3] argue that categorical systems like KL gradings are limited by incorrect assumptions that the progression of distinct OA features like JSN and osteophytes formation is linear and constant, and their relationships are proportional, and such grading systems are less sensitive to small changes in distinct features. Therefore, quantification of individual features of knee OA is required to overcome the problems with KL gradings and to improve the overall radiographic assessment of knee OA [1,3]. The osteoarthritis research society international (OARSI) has published a radiographic atlas of individual features to assess and to quantitatively evaluate the knee OA features [1].

Interactive methods like KOACAD [1] and KIDA [3] measure individual knee OA radiographic features such as joint space width (JSW), osteophyte area, sub-chondral bone density, joint angle, and tibial eminence height as continuous variables. These measurements were compared to KL gradings and significant differences were found between healthy knees and knees with OA. In this context, a trainable rule-based algorithm has also been proposed [53] to measure the minimum joint space width (mJSW) between the edges of the femoral condyle and the tibial plateau, and thus to monitor the progression of knee OA. Podsiadlo et al. [54] have used a slightly different approach for quantitative knee OA analysis. In this method, the trabecular bone regions of the tibia are automatically located as the ROI after delineating the cortical bone plates using active shape models, followed by fractal analysis of bone textures for the diagnosis of knee OA. In a similar approach, Lee et al. [56] use active shape models to detect the tibia and femur joint boundaries, and calculate anatomical geometric parameters to diagnose knee OA.

Even though these methods are simple to implement, objective, and accurate in evaluating radiographic knee OA, a great deal of manual intervention is required. Hence, these methods become very time-consuming and laborious when large numbers of subjects are to be investigated. Furthermore, the measurements from these methods are prone to inter- and intra-observer variability and in some cases they are subjective and not reproducible.

2.3.2 Automatic Classification

After the introduction of radiography-based semi quantitative scoring systems like KL gradings, the assessment of radiographic knee OA severity has been approached as an image classification problem [2, 15, 29–31]. According to the literature, the most common approach to classify knee OA images includes two steps: 1) extracting image features from the knee joints, 2) applying a classification algorithm on the extracted features. A brief review of such approaches is aras follows.

Subramoniam et al. [29,30] investigated two methods using: 1) the histograms of local binary pattern extracted from knee images and a k-Nearest neighbour classifier [29] and 2) Haralick features extracted from the ROI of knee images and a SVM [30]. Thomson et al. [15] proposed an automated method that uses features derived from tibia and femur bone shapes, and image textures extracted from the tibia with a simple weighted sum of the outputs of two random forest classifiers. Deokar et al. [31] investigated an artificial neural network based approach for knee OA images classification using grey level co-occurrence matrix (GLCM) textures, shape, and statistical features. Even though these methods claim high accuracy, the datasets are not publicly available and these datasets contain only a few hundred radiographs. The classification accuracies of all these methods for public datasets like the OAI and the MOST need to be studied to derive conclusive results.

In this context, there are two approaches in the literature that use large public datasets like OAI: 1) WNDCHRM[1], 2) an artificial neural network-based scoring system. Shamir et al. proposed WNDCHRM, a multi purpose medical image classifier to automatically assess knee OA severity in radiographs [2, 7]. A set of features based on polynomial decompositions, high contrast, pixel statistics, and textures are used in WNDCHRM. Besides extracting features from raw image pixels, features extracted from image transforms like Chebyshev, Chevbyshev-Fourier, Radon, and Gabor wavelets are included to expand the feature space [7, 9, 14]. From the entire feature space, highly informative features are selected by assigning feature weights based on a Fisher discriminant score for all the extracted features [2, 7, 14]. WNDCHRM uses a variant of the k-Nearest Neighbour classifier.

In a recent approach, Yoo et al. [32] have built a self-assessment scoring system and an artificial neural network (ANN) model for radiographic and symptomatic knee OA risk prediction. First, for developing a risk prediction model the association between risk factors and radiographic knee OA are investigated by multi variable logistic regression in this study. Next, ANNs are used to improve the performance of the scoring system. The prediction models are validated using two datasets: OAI[2] and KNHANES V-1[3]. The authors themselves have pointed out some limitations in this study. First, the study was based on a cross-sectional survey which had several defects due to medical views. For instance, the prevalence of disease was based on a health interview survey taken on one occasion. BMI, physical activity status, as well as knee pain could differ according to the time of measurement [32]. Second, the prediction models include knee pain as an important diagnostic criterion for symptomatic knee OA, which is subjective.

[1] Weighted Neighbour Distance using Compound Hierarchy of Algorithms Representing Morphology

[2] The osteoarthritis initiative

[3] Fifth Korean National Health and Nutrition Examination Survey

2.4 Discussion

According to the literature, the automatic quantification of knee OA severity involves two steps: 1) automatically detecting the ROI, 2) classifying the detected knee joints. Many previous studies investigated automatic methods for both localisation and classification of knee joint images, but still these tasks remain a challenge.

The common approaches in the literature for automatic detection of knee joints in radiographs include template matching [7, 55], active shape models and morphological operations [54], and a classifier-based sliding window method [46]. Template matching and active shape models based approaches do not generalise well and are slow for large datasets. Classifier-based methods that use hand-crafted features are subjective and the classification accuracy is influenced by the choice of extracted features. Therefore, there is still a need for an automated method for detecting knee joints in radiographs which gives high accuracy and precision. A deep learning based method for this is investigated in this thesis.

There are several approaches in the literature for knee OA image classification that have extracted and tested many image features such as Haralick textures [30], Gabor textures [15], GLCM textures [31], local binary patterns [29], shape, and statistical features of knee joints [31]. There is even an approach that uses a large set of features based on pixel statistics, object and edge statistics, texture, histograms, and transforms [4, 9, 14]. Different classifiers have been tested for knee OA images classification such as k-Nearest Neighbour [7, 29], SVM [30], and random forest classifiers [15]. However, all these approaches have achieved low multi-class classification accuracy and in particular classifying successive grade knee OA images still remains a challenging task. There is a need for a highly accurate real world automated system that can be used as a support system by clinicians and medical practitioners for knee OA diagnosis.

In recent years, many methods using manually designed or hand-crafted features have been outperformed by approaches that learn feature representations using deep neural networks. In particular, convolutional neural networks (CNN) have become highly successful in many computer vision tasks like object detection, face recognition, content based image retrieval, pose estimation, and shape recognition, and even in medical applications such as knee cartilage segmentation in MRI scans [24], brain tumour segmentation in magnetic resonance imaging (MRI) scans [25], multi-modality iso-intense infant brain image segmentation [26], pancreas segmentation in CT images [27], and neuronal membrane segmentation in electron microscopy images [28].

CNNs for automatically quantifying knee OA severity is investigated in this thesis. The next section introduces the necessary technical background and discusses the deep learning concepts and algorithms related to this thesis.

2.5 Public Knee OA Datasets

The data used for the experiments and analysis in this thesis are bilateral PA fixed flexion knee X-ray images. Figure 2.1 shows some samples of knee X-ray images from the dataset. Due to variations in X-ray imaging protocols, there are some visible artefacts in the X-ray images (Figure 2.1).

The datasets are from the Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST) in the University of California, San Francisco. These are standard public datasets used in knee osteoarthritis studies.

OAI Dataset

The baseline cohort of the OAI dataset contains MRI and X-ray images of 4,746 participants. In total 4,446 X-ray images are selected from the entire cohort based on the availability of KL grades for both knees as per the assessments by Boston University X-ray reading centre (BU). In total there are 8,892 knee images. Figure 2.2 shows the distribution as per the KL grades.

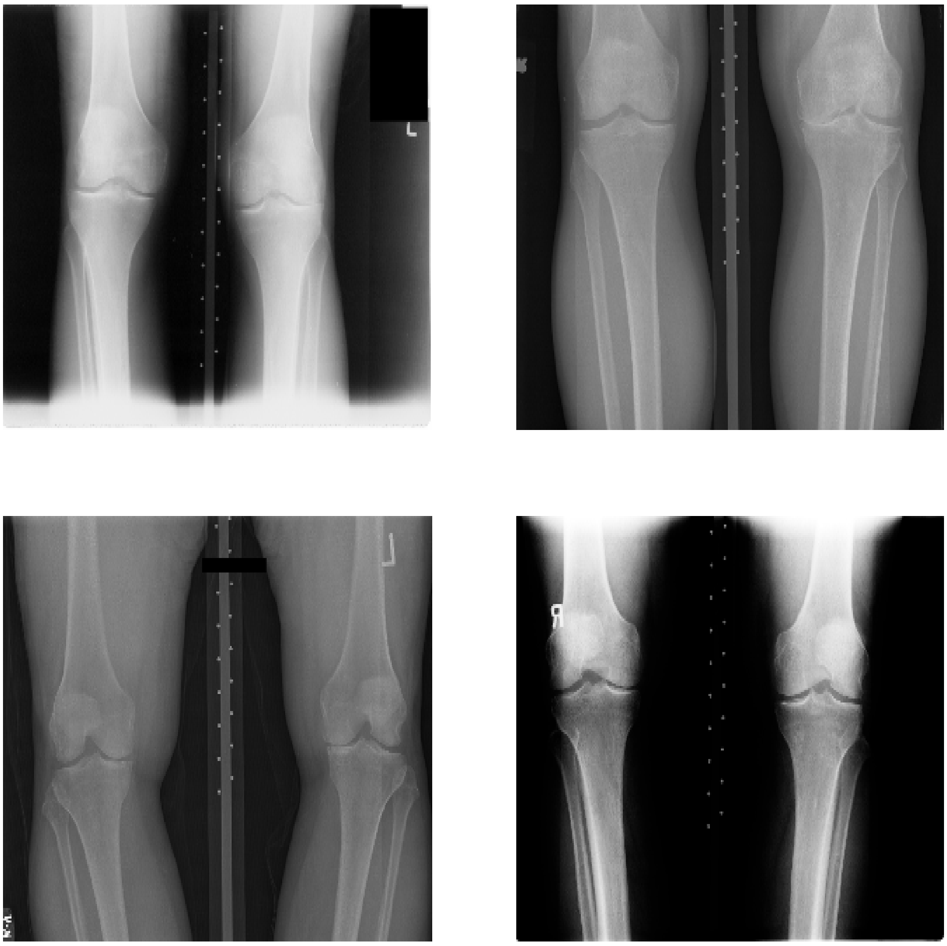


Figure 2.1: Samples of bilateral PA fixed flexion knee OA radiographs.

MOST Dataset

The MOST dataset includes lateral knee radiograph assessments of 3,026 participants. In total 2,920 radiographs are selected in this thesis based on the availability of KL grades for both knees as per baseline to 84-month longitudinal knee radiograph assessments. There are 5,840 knee images in this dataset. Figure 2.3 shows the distribution as per KL grades.

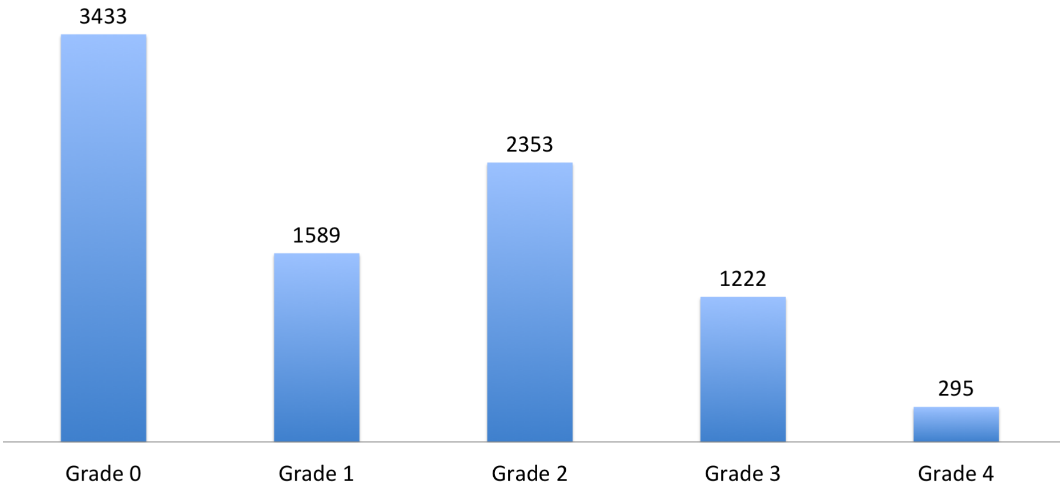


Figure 2.2: The OAI baseline data set distribution based on KL grades.

2.6 Deep Learning

There are many machine learning and signal processing techniques in the literature that use shallow architectures with one or two layers of non linear feature transformations [57]. Examples of shallow architectures include SVM, logistic regression, kernel regression, Gaussian mixture models (GMM), conditional random fields (CRF), multi-layer perceptron (MLP) with a single hidden layer, and maximum entropy models [57]. These architectures have been highly effective and have yielded promising results in several simple or well-constrained problems. However, their limited modelling and representational power may not be sufficient to deal with more complicated real-world applications. Deep learning algorithms and architectures can be effective and efficient for these applications.

Deep learning is composed of machine learning training algorithms to learn features using multi-layer networks with non linear processing units in each layer [57]. There are multiple levels of learned representations in deep learning, which attribute to various levels of abstraction; the levels correspond to hierarchical latent features and higher-level features are obtained from lower-level features [57]. Deep learning methods encompass deep neural networks, hierarchical

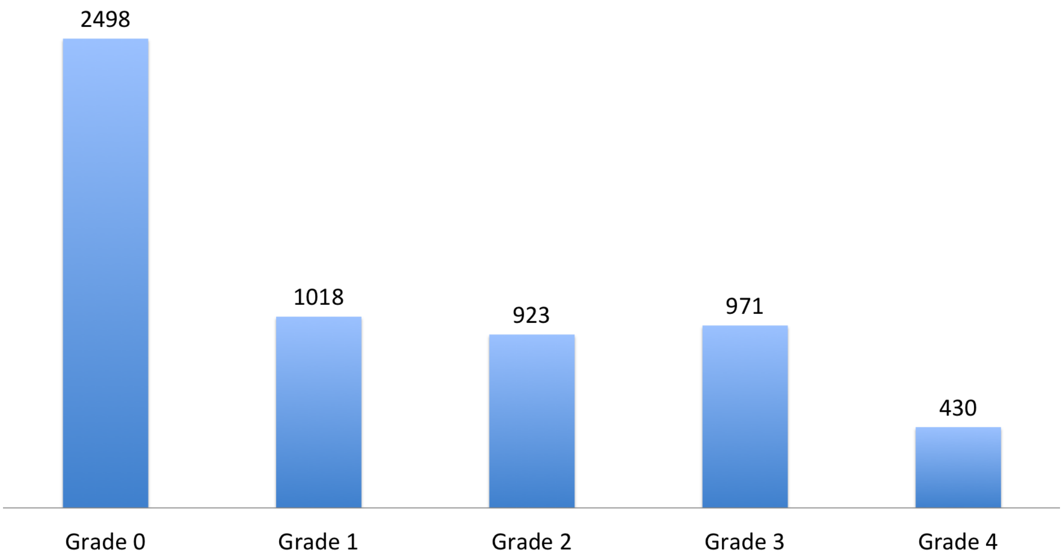


Figure 2.3: The MOST data set distribution based on KL grades.

probabilistic models, and various unsupervised and supervised feature learning algorithms.

Deep learning architectures are broadly classified into three classes: deep networks for unsupervised or generative learning, deep networks for supervised or discriminative learning, and hybrid deep architectures [57, 58]. Generative deep architectures characterise the high-order correlation properties of the observed or visible data for pattern analysis purposes and/or characterise the joint statistical distributions of the visible data and their associated classes. Discriminative deep architectures provide discriminative power for pattern classification, often by characterising the posterior distributions of classes conditioned on the visible data. The goal of hybrid architectures is discrimination but they benefit from the outcomes of generative architectures through better optimisation and/or regularisation [57, 58]. State-of-the-art deep learning architectures include multilayer deep neural networks, recurrent neural networks, convolutional neural networks (CNN), deep belief networks, deep boltzmann machines, stacked sparse auto encoders, and deep stacking networks. CNNs in particular have been tremendously successful in many real-world applications. In this thesis, the focus is on CNNs for automated knee OA assessment.

2.6.1 Convolutional Neural Networks

The visual cortex is responsible for processing visual information in the brain. The information is processed in a sequence of areas of brain in a low to high abstraction level [24]. The study of visual cortex shows that the neurons present get activated by stimuli generated by localised fields. Linear filtering in image processing is performed through convolution in the spatial domain (or element-wise multiplication in the frequency domain). However, the idea behind the CNN is to learn these filters in a data-driven manner. In machine learning, deep learning networks have multiple non-linear hidden layers and can represent the data in a hierarchical way with lower to higher abstraction. CNNs are a variant of the multilayer perceptron, which are inspired by the visual cortex and have deep architectures [24, 59].

Feed forward neural networks are briefly discussed before getting into the details of a CNN. A feed forward network in general has multiple layers. The first layer is the input and the last layer is the output. There can be one or more hidden layers in between the input and the output layers. Each layer is made up of neurons that have learnable weights and biases. The output of the hidden layer neurons is calculated as the weighted sum of all the input neurons, which is then passed through an activation function, which is often a non-linear activation function such as sigmoid or ReLU (rectified linear unit) [24]. There is a unique weight for each pair of neurons in the input layer and hidden layer, and they are connected to each other pairwise, and hence these layers are called as fully-connected layers. In multi layer perceptrons all the layers are fully connected, thus, the number of free parameters eventually becomes too large to handle and a large number of parameters can quickly lead to overfitting [59]. This is a problem for image data, where the number of inputs is large.

CNNs constitute a class of feed forward networks and they have a very similar architecture. CNNs are so-named due to the convolutional layers in their architectures. There are three main differences between a CNN and an ordinary feed forward network: local receptive fields, weight sharing, and spatial pooling or sub-sampling layers [59, 60].

Local Connectivity. The set of neurons in the preceding layer that affects the activation of a neuron is referred to as the neuron's local receptive field. This feature makes the CNNs well suited for learning effective representations from images, capturing the local substructure within the images [59]. The pixels that are close together in a image often tend to be strongly correlated while pixels that are far apart tend to be weakly correlated or uncorrelated. The CNN architecture captures this local structure within the image by constraining each neuron to depend only on a spatially local subset of the neurons in the preceding layer [60].

Shared weights. The other feature that distinguishes the CNN from the ordinary feed forward network is the fact that the weights in the network are shared across different neurons in the hidden layers. Sharing the weights across multiple neurons in a hidden layer translates to evaluating the same filter over multiple sub-windows of the input image. Each set of shared weights is called a kernel or a convolutional kernel. In this regard, the CNN can be viewed as effectively learning a set of filters, each of which is applied to all of the sub-windows within the input image. Using the same set of filters over the entire image forces the network to learn a general encoding or representation of the underlying data. Constraining the weights to be equal across different neurons also has a regularising effect on the CNN; in turn, this allows the network to generalise better in many visual recognition settings. The other advantage of weight sharing is that it substantially reduces the number of free parameters in the CNN, making it easier and more efficient to train [60].

Spatial pooling. Sub-sampling or spatial pooling is a form of non-linear down sampling. Spatial pooling serves two purposes: it reduces the dimensionality of the convolutional output and it provides a degree of translational invariance [24]. There are several non-linear functions to implement pooling and there are many pooling methods like sum, average, and max pooling, among which max pooling is most commonly used.

CNN Architectures

A CNN is comprised of more than one convolutional and sub-sampling layer(s), optionally followed by the fully connected layers like a standard multilayer neural network, and finally a softmax layer or regression layer to generate the desired outputs. CNNs exploit the 2-Dimensional structure of an input image to learn translation invariant features. This is achieved with local connections with shared weights followed by some form of pooling [57,59]. The main advantage of CNN over fully connected networks is that they are easier to train, they have fewer parameters with the same number of hidden units and they learn spatially invariant features [24]. Some examples of CNN architectures are LeNet [61], AlexNet [22], GoogleNet [62], VGG net [63], and ResNet [64].

2.6.2 Feature Learning

Feature learning refers to techniques that learn to transform raw data input to an effective representation for further higher-level processing such as classification, automatic detection, and segmentation. Feature learning approaches provide a natural way to capture cues by using a large number of code words (sparse coding) or neurons (deep networks), while traditional computer vision features, designed for basic-level category recognition, may eliminate many useful cues during feature extraction [65]. Deep neural networks are multi-layered and they are used to learn feature representations in the hidden layer(s). These representations are subsequently used for classification or regression at the output layer, and feature learning is an integral part of deep learning [16].

Manually designed or hand-engineered features often simplify machine learning tasks. Nevertheless, they have a few disadvantages. These features are often low-level as prior knowledge is hand-encoded and features in one domain do not always generalise to other domains [17]. In recent years, learning feature representations is preferred to hand-crafted features, particularly for fine-grained classification, because rich appearance and shape features are essential for describing subtle differences between categories [18].

The broad categories of feature learning are as follows.

(a) Supervised feature learning is attributed to learning features from data assigned with labels. Some of the state-of-the-art techniques that adopt supervised feature learning are deep neural networks (CNN, DBN), learning kernels (multiple kernel learning) [66], and multi-task learning [16].

(b) Semi-supervised feature learning uses unlabelled data to aid supervised learning. Some examples of semi-supervised learning: 1) label propagation [67], 2) DBNs with unsupervised pre-training and supervised fine-tuning [68, 69], 3) ladder networks [70], 4) student-teacher models [71], 5) learning features from videos or ego motion followed by supervised fine-tuning [72].

(c) Unsupervised feature learning is learning feature representations from data even without pre-assigned labels. The main purpose of unsupervised feature learning is to discover and capture the underlying structures from unlabelled input data, to detect and remove input redundancies and to preserve the essential aspects as useful features for classification [73]. The machine learning algorithms that use unsupervised feature learning include K-means clustering, Gaussian mixture model (GMM), principal component analysis (PCA), sparse coding, and auto-encoders [16]. Latent dirichlet allocation (LDA), a popular text processing algorithm [74] and co-occurrence statistics, a heuristic algorithm widely used in natural language processing [75], and the widely used visual bag of words based models are examples of unsupervised feature learning.

2.6.3 Transfer Learning and Fine Tuning

Training a CNN from scratch or full training is computationally expensive and requires a large amount of labelled training data. Creating a large annotated dataset in the medical domain is difficult where expert annotations are expensive and sometimes the diseases or lesions are scarce in the datasets [23]. Therefore, training deep neural networks from scratch particularly for medical applications can be challenging due to limited labelled medical data, and it demands a great deal of expertise for labelling the medical data. A promising alternative for training a CNN from scratch is fine-tuning a CNN pre-trained on a large labelled dataset (for instance ImageNet, which contains 1.2 million images with 1000 categories [22, 33]) from a different application domain using transfer learning [23, 76].

Training a CNN from a set of weights from a pre-trained CNN is referred to as fine-tuning. A common practice is to replace the last fully connected layer of the pre-trained CNN with a new fully connected layer whose number of nodes is equal to the number of classes in the desired application [76]. After initialising the weights of the last fully connected layer, either all layers or only a subset of layers at the top of the network are fine-tuned [23, 76]. The initial layers of a CNN in general learn generic features like edge detectors or colour blob detectors, which are applicable to many vision tasks, but the later layers learn progressively high-level features more specific to the classes of the targeted application. Figure 2.4 shows a CNN architecture: AlexNet [22] trained with the ImageNet dataset and some examples of the learned features. Thus, fine-tuning the last few layers of a CNN is sufficient for transfer learning. [23, 76].

2.7 Chapter Summary

In this chapter, the main concepts in the literature, the state-of-the-art and previous work related to assessing knee OA severity are reviewed. Also, the technical background: deep learning concepts and architectures related to this thesis are introduced.

According to the literature, the diagnostic pipeline in a computer aided

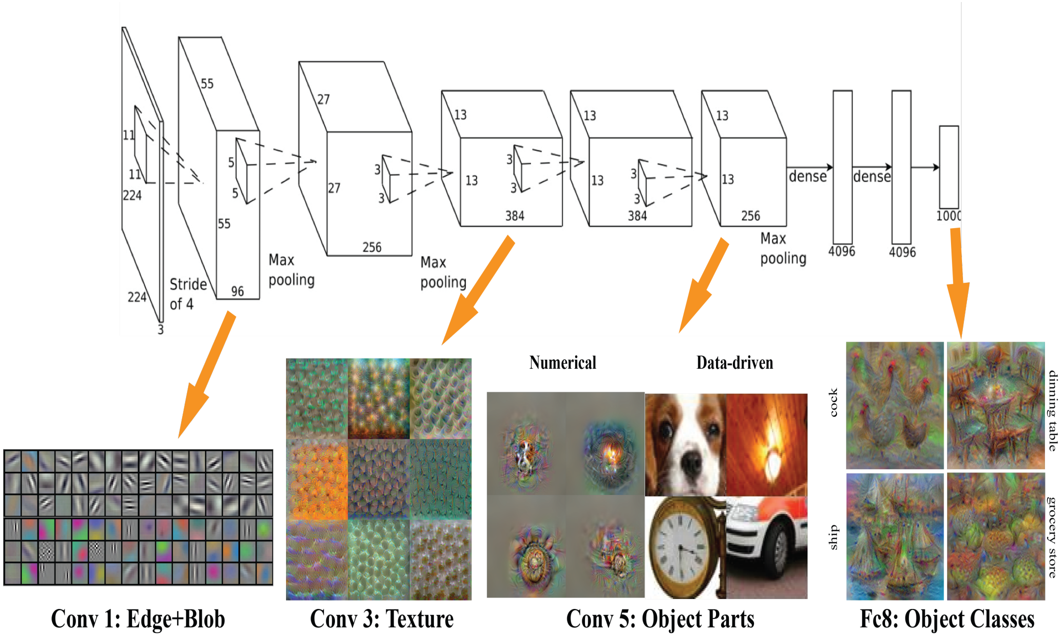


Figure 2.4: Some examples of the learned features from a CNN.

Source: `http://vision03.csail.mit.edu/cnn_art/`

assessment of knee OA consists of two steps: localising the knee joints and quantifying OA severity in the localised knee joints. First, the previous approaches in the literature for detecting the knee joints in radiographs are classified into manual, semi-automatic or interactive, and automatic methods. Manual methods are subjective, highly reliant on experience and time-consuming when large number of subjects are to be examined. Interactive methods sometimes lack reproducibility and they are prone to inter- and intra-observer variations. These drawbacks are overcome in automatic methods. There have been several attempts in the literature to automatically detect knee joints in radiographs. However, this task still remains a challenge. There is a need for a highly accurate automated method for this task.

There are two common approaches in the literature for assessing radiographic knee OA severity: 1) automatic classification of knee joint images based on KL grades 2) interactive methods that quantify the distinct pathological features of knee OA such as joint space narrowing (JSN), osteophytes area and sclerosis. Even though the interactive methods are objective and accurate, a great deal of manual intervention is required and these methods may become laborious and time-consuming for a large number of investigations. Some of these drawbacks are overcome in automatic methods, however, these approaches achieve low multi-class classification accuracy and classifying successive grade knee OA images still remains a challenge.

The next chapter presents the baseline approaches for automatically localising the knee joints and automatically quantifying OA severity on the localised knee joints.

Chapter 3

Baseline Methods

3.1 Introduction

This chapter presents the baseline approaches and experiments for automatically quantifying the knee OA severity from the X-ray images that will be used as the basis for comparison in this thesis. The automatic assessment of knee OA mainly involves two steps: 1) automatically detecting and extracting the region of interest (ROI) for localising the knee joints in the X-ray images, 2) classifying the localised knee joints based on the Kellgren & Lawrence (KL) grades. The objective of this chapter is to create strong baseline methods based on the existing state-of-the-art so that later more complex approaches can be compared with these methods.

The assessment of knee osteoarthritis (OA) severity has traditionally been approached as an image classification problem [2, 9, 32]. Shamir et al. [9, 34] proposed WNDCHRM, a multi purpose medical image classifier for classifying knee OA images using radiographs based on the KL grades and reported promising results for detecting knee OA in the minimal, moderate, and severe stages. WNDCHRM uses several hand-crafted features and a weighted nearest neighbour classifier to classify the knee OA images. In this chapter, the state-of-the-art hand-crafted features and classification methods to automatically assess knee OA severity are investigated, and the objective is to improve the overall classification accuracy.

First, the hand-crafted features that are successful in other computer vision tasks such as histogram of oriented gradients, [77], local binary patterns [78], and sobel gradients [79] are tested for knee OA images classification. Also, conventional classifiers such as k-nearest neighbour (kNN) classifier, support vector machine (SVM), and support vector regression (SVR) are tested for classifying the knee OA images. The feature space is expanded by selecting highly influential features in WNDCHRM based on feature ranking, such as tamura and haralick texture features, gabor textures, and zernike features.

Next, automatic feature learning is investigated in an attempt to improve the classification of the knee OA images. Inspired by the success of convolutional neural networks (CNN) in many computer vision tasks, the use of CNN features for the classification of the knee OA images is proposed in this thesis. Initially, the well-known and widely-used VGG ILSVRC 16 layer network is selected to extract features. The features are extracted from two different layers: the final pooling and the final fully-connected layer of this network and a SVM for classification is used.

Shamir et al. [2] proposed template matching for automatically detecting and extracting the knee joints from the radiographs. In this approach, the centre of the knee joints are detected and used as a reference to extract a fixed size region from the radiographs around the centre. This method is implemented as a baseline. Next, to improve the localisation of the knee joints, a SVM-based method using Sobel horizontal image gradients as features is proposed to detect the centre of the knee joints.

The remainder of this chapter is organised as follows: Section 3.2 introduces the dataset used for the experiments. Section 3.3 presents two methods for the automatic detection and localisation of the knee joints: 1) template matching and 2) SVM classification with Sobel horizontal gradients. Section 3.4 presents two approaches for classification of the knee joints using: 1) hand-crafted features, 2) CNN features. The outcome of these methods is compared to WNDCHRM classification. Section 3.5 summarises the baseline methods for localisation and classification of knee OA images, and presents the conclusions.

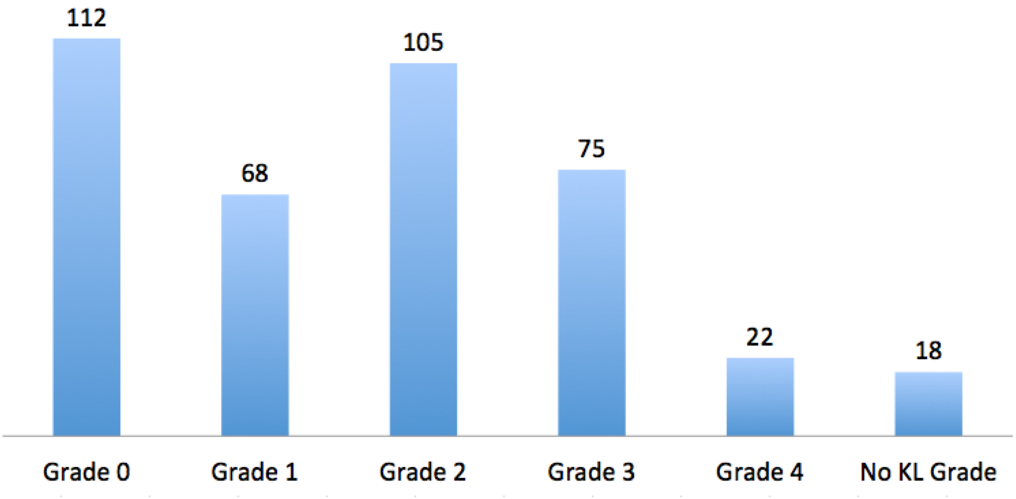


Figure 3.1: The OAI baseline data (200 X-ray images) distribution as per KL grades.

3.2 Dataset

The dataset used for the initial experiments is the knee X-ray images from the baseline data sample of 200 progression and incidence cohort subjects under the knee OA study. In total, 191 radiographs (382 knee joint images) have the assigned Kellgren & Lawrence (KL) grades. Figure 3.1 shows the distribution of the images as per KL grades (grade 0 to 4). This is a relatively small data set containing only X-ray images and MRI of 200 subjects. After the initial experiments, the entire dataset with 4,796 subjects from the Osteoarthritis Initiative (OAI) was acquired. As a preprocessing step, histogram equalisation is performed on all the X-ray images for intensity level normalisation. Figure 3.2 shows a few samples of X-ray images before and after histogram equalisation.

To investigate the classification of the knee OA images independent from the localisation of the ROI; the knee joint regions are manually cropped from the radiographs and resized to 200×300 pixels. The knee joint images are flipped left–right to generate more training data.

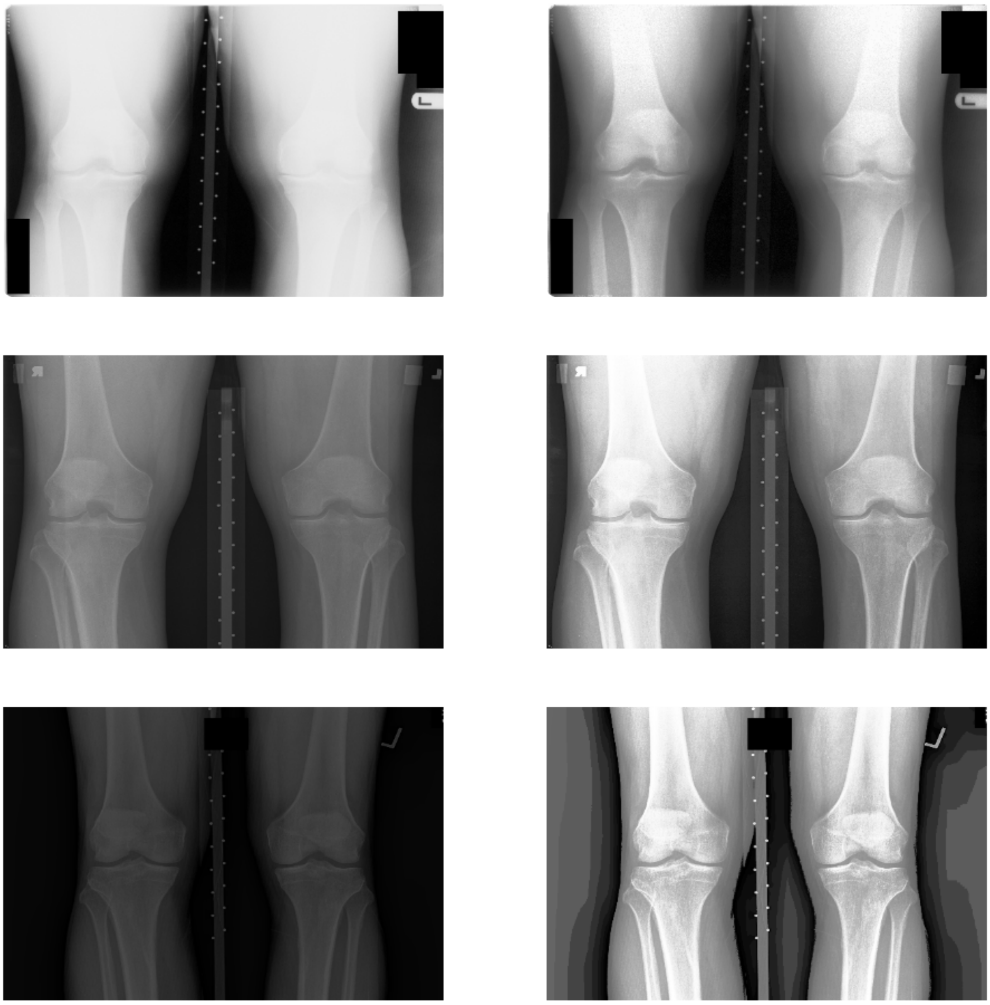


Figure 3.2: Samples of X-ray images before (left) and after (right) histogram equalisation.

3.3 Automatic Detection of Knee Joints

Classification of knee OA images and the assessment of severity conditions can be achieved by examining the characteristic features of knee OA: variations in the joint space width and the osteophytes (bone spurs) formations in the knee joints [1]. Radiologists and medical practitioners examine only the knee joint regions in the X-ray images to assess knee OA. Hence, the region of interest (ROI) for classifying knee OA images is only the knee joint regions (left and right knees). Figure 3.3 shows the ROI in a X-ray image. The author believes that it is better to focus

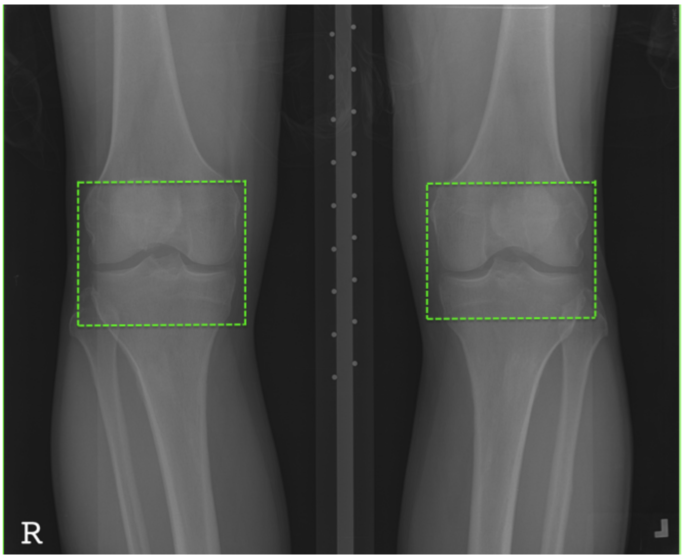


Figure 3.3: A knee OA X-ray image with the region of interest: the knee joints.

on the ROI instead of the entire X-ray image for accurate classification and this is also computationally economical. For these reasons, automatically detecting and extracting the knee joint regions from the X-ray images becomes an essential preprocessing step, before classification.

As a baseline, template matching for the automatic detection of the knee joints is implemented. In the following section, the implementation details and outcomes of this method are discussed.

3.3.1 Template Matching

In digital image processing, template matching is a technique for finding portions of an image that are similar to a standard template image. Shamir et al. [2] proposed this approach for automatically detecting the centre of the knee joints. As a baseline, the template matching approach is adapted. The steps involved in this method are as follows:

- First, the radiographs are downscaled to 10% of the original size and subjected to histogram equalisation for intensity normalisation. This step is followed as proposed by Shamir et al. [2].

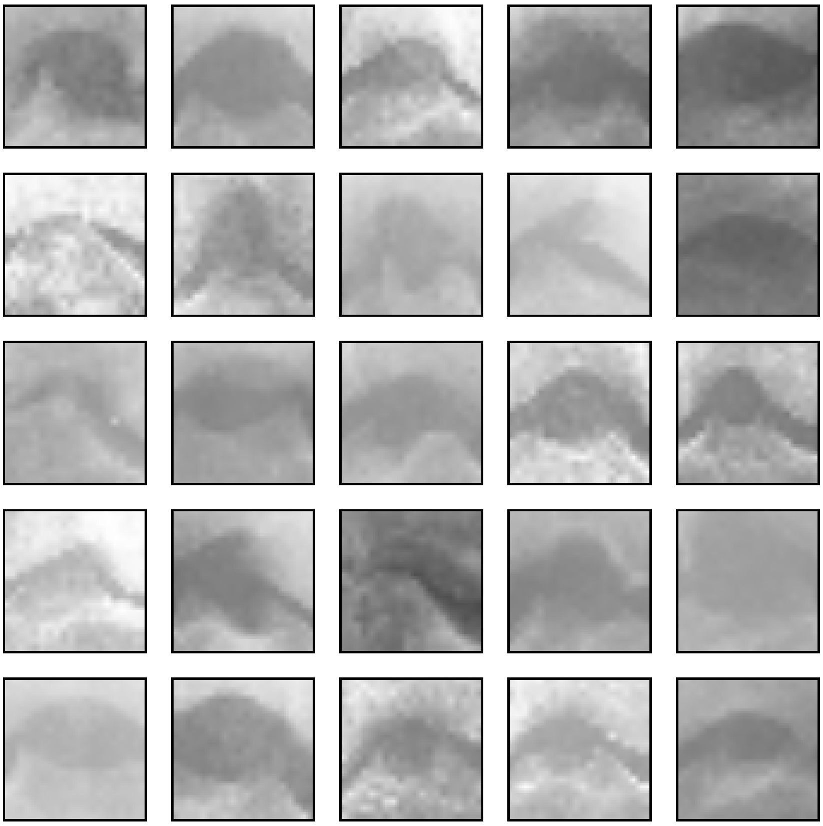


Figure 3.4: Pre-selected knee joint centres (20×20 pixels) extracted from knee joint images for template matching.

- An image patch (20×20 pixels) containing the centre of the knee joint is taken as a template. 5 image patches are taken from each grade, so that in total 25 patches are pre-selected as templates. Figure 3.4 shows the pre-selected knee joint centres of size 20×20 pixels extracted from the knee joint images as templates.

- Each image is scanned by an overlapping (20×20) sliding window. For each location at an interval of 10 pixels, distances (Euclidean) between an image patch (20×20 pixels) and 25 pre-selected templates (patches with knee joint centre) are computed using;

$$dist_{i,w} = \sqrt{\sum_{y=1}^{20}\sum_{x=1}^{20}(I_{x,y} - W_{x,y})^2}$$

, where $I_{x,y}$ is the intensity of pixel (x, y) in the knee joint image I, $W_{x,y}$ is the intensity of pixel (x, y) in the sliding window, and $dist_{i,w}$ is the Euclidean distance between the knee joint image (I) and the sliding window W.

- In total, 25 different distances are calculated at each location of the sliding window for the 25 templates, and the shortest among the 25 distances is recorded.

- The window with the smallest Euclidean distance is selected as the centre of the knee joint after scanning the image with a sliding window and a fixed size region (700×500 pixels) around this centre is extracted as the knee joint region from the X-ray image.

- The input X-ray images are horizontally split in half to isolate left and right knees separately and the sliding window is run on both halves.

Experiments and Results

For the experiments on template matching, the baseline data sample of 200 progression and incidence cohort subjects under the knee OA study is used. This dataset contains in total 191 X-ray images (382 knee joints) and it is a subset of the large OAI dataset.

In this implementation, five different sets of templates (each set with 25 templates) are used to show the influence of templates on knee joint detections. The templates are selected from a separate training set. visual inspection is used to evaluate the results of template matching by plotting a bounding box (20×20 pixels) on the image patch that recorded the shortest Euclidean distance after template matching. Table 3.1 shows the total number of true positives: the detected knee joint centres, the total number of false positives and the precision.

It is clearly evident from the results (Table 3.1) that template matching is not precise in detecting the knee joints and that the detections are heavily dependent on the choice of templates. The number of templates is increased to 50, but there is no

Table 3.1: Detection of knee joint centres using template matching method.

Templates	True Positives	False Positives	Precision
Set 1	87	295	22.8 %
Set 2	78	304	20.4 %
Set 3	99	283	25.9 %
Set 4	**116**	**266**	**30.3 %**
Set 5	55	327	14.4 %

further improvement in the results. The reason for low-performance of the template matching is that the computations are mainly based on the intensity level difference of an image patch and a template, and there are possibilities for image patches not around the knee joint, having the shortest Euclidean distance to a template in the set and thus, being detected as matches. In the next section, a new SVM-based method is investigated to improve the detection of the knee joints.

3.3.2 SVM-based Detection

Standard template matching is not scalable and produces poor detection accuracy on large datasets like the OAI. A classifier-based model to automatically detect the knee joints in the X-ray images is proposed in this thesis. The idea is to use well-known Sobel edge detection [79] for detecting the knee joints. The two major steps involved in this method are 1) training a classifier and 2) developing a sliding window detector.

Training a Classifier

First, image patches (20×20 pixels) are generated from the input X-ray images. The image patches containing the knee joint centre (20×20 pixels) are used as positive samples and randomly sampled patches excluding the knee joint centre are used as negative samples. In total, 200 positive and 600 negative samples are used. The image patches (samples) are split into training (70%) and test (30%) sets. Sobel horizontal image gradients are extracted as features from all these samples to train a classifier. The powerful and well-known SVM is used for classification. A linear SVM is fitted with default parameters: C=1, and linear kernel, using Sobel horizontal image gradients as the features.

Before settling on Sobel horizontal image gradients as features, the state-of-the-art features such as histogram of oriented gradients, Tamura and Haralick textures, and the Gabor features were tested. The HOG features are highly accurate and efficient in object detection and human detection [77]. The Tamura and Haralick textures, and Gabor features are highly influential and top-ranked among the features used in WNDCHRM for knee OA image classification [1, 2, 7, 14]. The Sobel operator or Sobel filter uses vertical and horizontal image gradients to emphasise the edges in images [79]. From these, the horizontal image gradients are used as the features for detecting the knee joints centres. Intuitively, the knee joint images primarily contain horizontal edges that are easy to detect.

Sliding Window Detector

To detect the knee joint centre from both left and right knees, input images are split in half to isolate left and right knees separately. A sliding window (20×20 pixels) is used on either half of the image, and the Sobel horizontal gradient features are extracted for every image patch. The image patch with the maximum score based on the SVM decision function is recorded as the detected knee joint centre, and the area (200×300 pixels) around the knee joint centre is extracted from the input images using the corresponding recorded coordinates. Figure 3.5 shows an instance of a detected knee joint and the extracted ROI in a X-ray image.

Results and Discussion

In total, 200 image patches with the knee joint centres as positive samples and 600 image patches that exclude the centre of knee joint as negative samples are used. These images are split into training (70%) and test (30%) sets. Fitting a linear SVM with the training data produced a 5-fold cross validation accuracy of **95.2%**

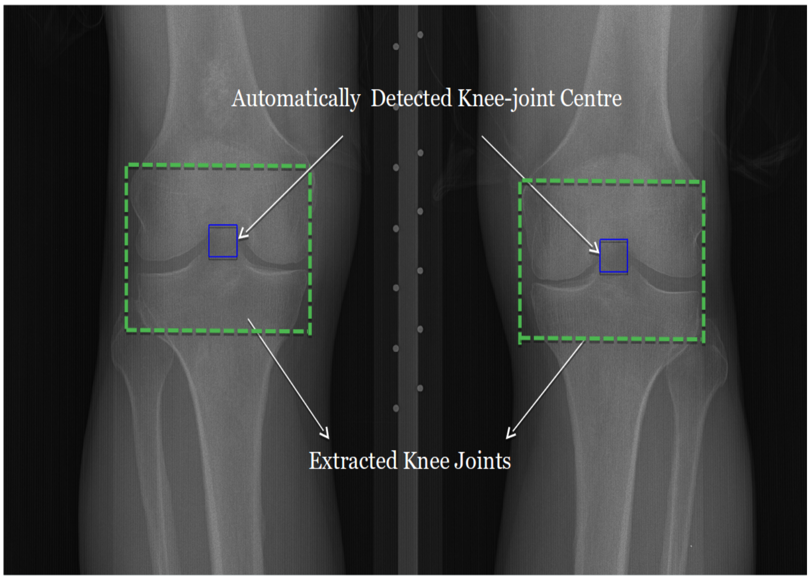


Figure 3.5: Detecting the knee joint centres and extracting the knee joints.

and an accuracy of **94.2%** for the test data. Table 3.2 shows the precision, recall, and F_1 scores of this classification. To evaluate the automatic detection, the ground truth is generated by manually annotating the knee joint centres (20×20 pixels) in 4,446 radiographs using an annotation tool that we developed, which recorded the bounding box (20×20 pixels) coordinates of each annotation.

Table 3.2: Classification metrics of the SVM for detection.

Class	Precision	Recall	F_1score
Positive	0.93	0.84	0.88
Negative	0.95	0.98	0.96
Mean	0.94	0.94	0.94

The well-known Jaccard index (JI) is used to give a matching score for each detected instance. The Jaccard index JI(A,D) is given by,

$$JI(A, D) = \frac{A \cap D}{A \cup D} \tag{3.1}$$

where A, is the manually annotated and D is the automatically detected knee joint centre using the proposed method.

Table 3.3: Comparison of template matching and the proposed SVM-based method.

Method	$JI = 1$	$JI \geq 0.5$	$JI > 0$
Template Matching	0.3 %	8.3 %	54.4 %
Proposed Method	1.1 %	38.6 %	**81.8 %**

Table 3.3 shows the resulting average detection accuracies based on thresholding of Jaccard indices. The mean JI for the template matching and the classifier methods are **0.1** and **0.36**. From Table 3.3, it is evident that the proposed method is more accurate than template matching. This is due to the fact that template matching relies upon the intensity level difference across an input image. Thus, it is prone to matching a patch with small Euclidean distance that does not actually correspond to the knee joint centre. Also, the templates are varied in a set, and it is observed that the detection is highly dependent on the choice of templates. Template matching is similar to a k-nearest neighbour classifier with $k = 1$.

The reason for higher accuracy in the proposed method is the use of horizontal edge detection instead of intensity level differences. The knee joints primarily contain horizontal edges and thus are easily detected by the classifier using horizontal image gradients as features. The proposed method is approximately 80× faster than template matching; for detecting all the knee joints in the dataset comprising $4,446$ radiographs, the proposed method took $\sim$9 minutes and the template matching method took $\sim$798 minutes.

Despite sizeable improvements in accuracy and speed using the proposed approach, detection accuracy still falls short. Therefore the manual annotations are used to investigate KL grade classification performance independently of knee joint detection.

3.4 Classifying Knee OA Images

Previous work on automatic assessment of radiographic knee OA has used WNDCHRM, the multi purpose medical image classifier [2, 9, 14]. High classification accuracies (80% to 91%) have been reported using WNDCHARM for classifying the extreme stages of knee OA: grade 0 (normal) vs grade 4 (severe), grade 0 vs grade 3 (moderate). However, the classification accuracies of images belonging to successive grades are low (55% to 65%). The overall classification accuracy of knee OA needs improvement for real-world computer aided diagnosis [1,2,7].

As a baseline, the hand-crafted features that have been successful in other computer vision tasks such as histogram of oriented gradients [77], local binary patterns [78], and Sobel gradients [79] are investigated and that are not included in the previous studies to assess knee OA severity. Conventional classifiers such as k-nearest neighbour classifier, SVM, and support vector regression (SVR) are used for classification. Next, in an attempt to improve the classification accuracy, the CNN features extracted from the pre-trained VGG-16 layer network are used. To benchmark the results from these approaches, the WNDCHRM classification results are used.

3.4.1 WNDCHRM Classification

WNDCHRM is an open source utility for biological image analysis and medical image classification [2,9,34]. In WNDCHRM, a generic set of image features based on pixel statistics (multi–scale histograms, first four moments), textures (Haralick and Tamura features), factors from polynomial decomposition (Zernike polynomials), and transforms (Radon, Chebyshev statistics, Chebyshev-Fourier statistics) are extracted. For feature selection, every feature is assigned a *Fisher score*[1] and 85% of the features with lowest *Fisher scores* are rejected and the remaining 15% of the features are used for classification [2].

[1] Fisher score is one of the widely used method for determining the most relevant features for classification

Table 3.4: Results of WNDCHRM Classification.

Classification	Grades	Accuracy
Binary	G0 vs G1	66.7 %
	G1 vs G2	48.3 %
	G2 vs G3	60 %
	G3 vs G4	55 %
	G0 vs G2	48.3 %
	G0 vs G3	70 %
Multi-class	G0 to G4	28.3 %
	G0 to G3	35.8 %

Experiments

The dataset used for the initial experiments to classify knee OA images using WNDCHRM are taken from the baseline data sample of 200 progression and incidence cohort. After histogram equalisation and mean normalisation of the X-ray images, the knee joints are extracted manually from the radiographs. The extracted knee joints are split into training (70%) and test (30%) sets. The WNDCHRM command line program is used to classify the extracted knee joint images. WNDCHRM uses a variant of k-nearest neighbour classifier.

Results and Discussion

The baseline dataset is not balanced and there are only 44 samples available in KL grade 4. Figure 3.1 in Section 3.2 (Page 41) shows the distribution of the entire data set. Given the limited number of images in this class, only a small number of images are used for training and testing (35 images for training and 9 images for testing) for multi-class classification. For other classifications 100 images are used for training and 30 images for testing.

It is evident from the results (Table 3.4) that the multi-class classification accuracy and successive grades classification accuracies are very low. The reason for low classification accuracy is that the features used for classification are not capable of capturing the minute structural and morphological variations in the

Table 3.5: Classification results of the proposed methods using hand-crafted features.

Grades	WNDCHRM	SVM classification with hand-crafted features			
		HOG	LBP	Sobel	Combining all
G0 vs G1	**66.7 %**	53.3 %	58.3 %	58.3 %	55 %
G1 vs G2	48.3 %	48.3 %	53.3 %	**58.3 %**	51.6 %
G2 vs G3	60 %	60 %	60 %	56.7 %	**63.3 %**
G3 vs G4	55 %	**65 %**	**65 %**	50 %	**65 %**

knee joints between the successive grades. Next, the state-of-the-art hand-crafted features are investigated in an attempt to improve the classification accuracy.

3.4.2 Classification using Hand-crafted Features

Histogram of oriented gradients, local binary patterns, and Sobel Gradients are tested for classifying knee OA images [2, 7, 14]. These features are not used in the previous studies. HOG describes the local object shape and appearance within an image by the distribution of intensity gradients or edge directions and the HOG descriptor was successful in human detection [77]. LBP is powerful for image texture classification. LBP uses local spatial patterns and grey scale contrast as measures for texture classification [78].

Experiments with HOG, LBP and Sobel descriptors

Once again the images from the baseline data sample of 200 progression and incidence cohort is used. The HOG, LBP and Sobel descriptors are extracted from the knee joint images and a SVM is used for classification. Table 3.5 shows the classification results of successive grades of knee OA images using a SVM and the feature space included the HOG, LBP and Sobel gradients.

There is no large improvement in the classification accuracies using HOG, LBP and Sobel gradients features with SVM classification from the previous results with the WNDCHRM classification. To improve the classification, thus the features space is expanded by including highly effective and top-ranked features from the

Table 3.6: Classification results of WNDCHRM and the proposed methods using hand-crafted features.

Grades	WNDCHRM	Proposed Methods		
		kNN	SVM	SVR
G0 vs G1	**66.7%**	55%	60%	60%
G1 vs G2	48.3%	**61.7 %**	46.7%	48.3%
G2 vs G3	**60%**	51.7%	55%	**60%**
G3 vs G4	**55%**	35%	50%	45%
G0 vs G2	48.3%	46.7%	55%	**56.7%**
G0 vs G3	**70%**	48.3%	58.3%	60%

WNDCHRM classification.

Expanding the feature space

The features based on pixel statistics and textures such as Tamura, Haralick, Gabor and Zernike are used for classification. These features are used in the WNDCHRM classification. Tamura texture features represent contrast, coarseness and directionality of an image [80]. Haralick features are the statistics computed on the co-occurrence matrix of an image [81]. Gabor textures are based on Gabor wavelets and the image descriptors are computed using Gabor transform of an image [82]. Zernike features are obtained by the Zernike polynomial approximation of an image [83]. The feature space for classification is formed by simple concatenation of all the extracted features into a super vector following the early fusion approach.

Classification Results and Discussion

First, a SVM is used with the extracted features for classifying knee OA images. Next, a k-nearest neighbour classifier and support vector regression (SVR) are tested for classification. In total, 100 knee joint images are taken for training and 30 for test set in each grade. Table 3.6 shows the classification accuracy of the WNDCHRM classifier and the classification using kNN, SVM, and SVR.

When comparing the classification results of the proposed methods (SVM, kNN, and SVR) to the WNDCHRM classification, for some cases the results are slightly better and promising. Nevertheless, there is a need for a more significant improvement in the classification results. In these experiments, a subset of features from WNDCHRM such as Tamura & Haralick texture features, Gabor wavelet features, and Zernike features were extracted and used for classification. In addition to these features HOG, LBP, and Sobel Gradients were tested. It was found that by further expanding the feature space by including features from WNDCHRM based on transforms such as Radon, Chebyshev, FFT, and Wavelet, and compound image transforms such as Chebyshev-FFT, Chebyshev-Wavelet, and Wavelet-FFT classification can be improved. However, the author believes that learning feature representations can be more effective for fine-grained knee OA classification. In the following section, the state-of-the-art CNN features are investigated for classifying knee OA images.

3.4.3 Classification using CNN Features

In many recent computer vision tasks and medical applications, CNNs have been shown to outperform existing approaches that use hand-crafted features [20, 21]. Prasoon et al. have successfully developed and trained a triplanar CNN from scratch for segmenting the articular cartilage in knee MRIs [24]. There are previous works in the literature wherein a pre-trained CNN has been successfully adapted to the target application. For instance, Chen et al. [84] proposed the use of a pre-trained CNN for localising standard planes in ultrasound images. Carneiro et al. [85] have shown promising results for classification of unregistered multi view mammograms using a pre-trained CNN. Shin et al. [86] used pre-trained CNNs after fine-tuning to automatically map medical images to document-level topics and sub-topics. Gao et al. [87] fine-tuned all layers of a pre-trained CNN for automatic classification of interstitial lung diseases. Motivated by these approaches, the use of CNN features for classifying knee OA images is investigated.

As a baseline, the VGG-16 layers network [63] is investigated. Figure 3.6 shows the architecture of the VGG-16 layers network. This network was developed and trained by the visual geometry group (VGG) from the University of Oxford. This network achieved outstanding performance on the classification and localisation tasks in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. The VGG-16 network is pre-trained on the ImageNet dataset [63]. The ImageNet dataset contains more than 1 million annotated images for 1000 classes [33].

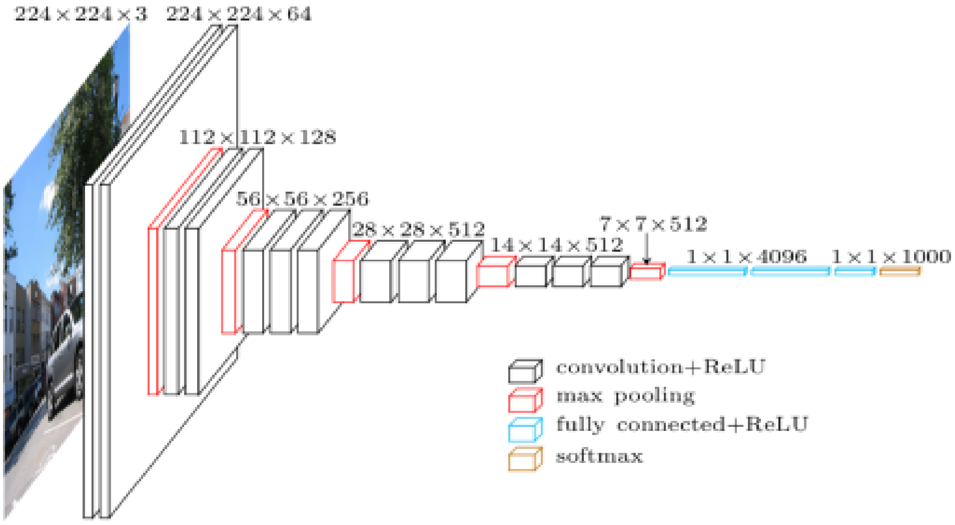


Figure 3.6: The VGG-16 layers network architecture.

Source: `https://www.cs.toronto.edu/{\sim}frossard/post/vgg16/`

Dataset

The same dataset used in the previous experiments i.e. the OAI baseline dataset, is used to train and test the CNN. The same training (70%) and test (30%) split is used. The preprocessing steps include manually segmenting the knee joint regions and resizing to 224×224 pixels, as per the input requirement of the VGG-16 network.

CNN Features

Features are extracted from the different pooling and fully-connected layers of the VGG-16 network. The results for the pooling layer (pool5) and the last fully

Table 3.7: Classification results of WNDCHRM and the proposed methods.

Grades	WNDCHRM	Hand-Crafted Features			CNN Features	
		kNN	SVM	SVR	Pool5	FC7
G0 vs G1	**66.7 %**	55 %	60 %	60 %	65 %	65 %
G1 vs G2	48.3 %	**61.7 %**	46.7 %	48.3 %	40 %	35 %
G2 vs G3	60 %	51.7 %	55 %	60 %	**63.3 %**	58.3 %
G3 vs G4	50 %	35 %	50 %	45 %	85 %	**90 %**
G0 vs G2	48.3 %	46.7 %	48.3 %	**56.7 %**	51.6 %	48.3 %
G0 vs G3	70 %	48.3 %	58.3 %	60 %	**76.6 %**	73.3 %

connected layer (fc7) are shown. The dimension of the features from the pool5 layer are 7×7×512 for each image. The pool5 features are flattened into a row vector for each image to form the feature space for classification. The dimension of the fc7 features are 4,096. A Liblinear SVM with the CNN features is used to classify the knee OA images. These results are compared to the classification results obtained from WNDCHRM, and the proposed methods using hand-crafted features with kNN classifier, SVM, and Support Vector Regression.

Classification Results

Table 3.7 shows the binary classification accuracies of knee OA images using WNDCHRM and the proposed methods using the hand-crafted features and the CNN features. These results show slight improvements in the successive grades classification using CNN features in comparison to the WNDCHRM and the proposed methods.

Table 3.8 shows the multi-class classification of knee OA images using WNDCHRM and the proposed method using the CNN features and a SVM. The classification accuracies obtained using pool5 features of VGG-16 network shows improvement over the WNDCHRM.

The VGG-16 layer CNN has been pre-trained and hyper parameters are tuned for general images in ImageNet dataset. Though this network is not trained for medical images, using the same network for knee OA image classification has yielded better

Table 3.8: Multi-class classification using WNDCHRM and CNN features.

Grades	WNDCHRM	CNN Features	
		Pool5	FC7
G0 to G3	35.8 %	**36.6 %**	33.3 %
G0 to G4	28.3 %	**33.8 %**	30.7 %

results than expected, despite using such a small dataset (100 images from each grade) for training. Clearly, the use of CNN for knee OA images classification is promising avenue to explore. The classification results can be further improved by tuning the hyper parameters and training the CNN with more image samples. These results are shown in Section 5.2.

3.5 Chapter Summary

As a baseline for the automatic detection of the knee joints, template matching was implemented. For this method, the precision of detection is low (∼30%), as the computations are mainly based on intensity-level differences. Therefore, a SVM based method is proposed in this thesis to improve the automatic detection of knee joints. This method gives better results with a precision of detection above 80%. The rationale behind better results in the proposed method is the use of Sobel horizontal image gradients and performing horizontal edge based discrimination compared to simple intensity-level discrimination in the template matching method.

Two approaches were investigated for classifying knee OA images using hand-crafted features and the learned features in a supervised manner using a CNN. The first approach used hand-crafted features with conventional classifiers such as SVM, kNN classifier, and SVR. The results obtained with SVM, kNN, and SVR are promising and can be further improved by including more features based on transforms and compound image transforms. However, the author believes that learning feature representations can be more effective than testing more hand-crafted features. The second approach is based on supervised feature learning using the

VGG-16 network. The classification of knee OA images using a pre-trained CNN gives promising results.

In this chapter, the baseline approaches and the initial experiments to automatically detect knee joints and to classify knee joints are reported. The outcomes of the proposed methods are promising and motivates the use of deep learning for automatic localisation of knee joints and classification of the localised knee joints to automatically quantify knee OA severity. As a result, these are the main focus and subject of the forthcoming chapters. The next chapter describes the proposed methods for automatically localising the knee joint region using fully convolutional neural networks.

Chapter 4

Automatic Localisation of Knee Joints

4.1 Introduction

Detecting and extracting the knee joints from X-ray images is an essential step before classifying the knee OA images and for large datasets automatic methods are preferable. Shamir et al. proposed template matching for the automatic detection of the centre of knee joints [2, 7] and then a fixed size region with reference to the detected centre is extracted as the region of interest (ROI). Standard template matching produces poor detection accuracy on large datasets like OAI dataset. To improve this, a linear SVM is fitted with the Sobel horizontal image gradients as features to detect the knee joints. Though this method gives a sizeable improvement in the detection accuracy it still falls short of perfect detections. These methods are presented in the previous chapter. In this chapter, the focus is on fully convolutional networks (FCN) to automatically localise the knee joints in X-ray images.

A typical CNN architecture consists of three main types of layers: convolutional, pooling and fully-connected or dense layers. A FCN is similar to a CNN, but the fully-connected layers are replaced by convolutional layers [88]. A FCN consists of mostly convolutional layers and if pooling layers are used, then suitable up-sampling layers are added before the last convolutional layer. The two major differences of FCNs over CNNs can be summarised as:

- FCNs are trained end-to-end to make pixel-wise predictions [88]. Even the decision-making layers at the last stage of the network use learned convolutional filters.
- The input image size need not be fixed as there are no fully-connected layers in the FCN. CNNs with fully connected layers can operate only on a fixed size input.

FCNs have achieved great success in semantic segmentations of general images [88]. Recent approaches using FCNs for medical image segmentation show promising results [89–91]. Motivated by this, the use of FCN is investigated in this chapter for automatically detecting the knee joints. Two approaches are developed for localising the knee joints: 1) training a FCN to detect the centre of knee joints and extract a fixed-size region around the detected centre, 2) training a FCN to detect the ROI and thus extract the knee joints directly.

The remainder of this chapter is structured as follows: Section 4.2 introduces the localisation of the knee joints with reference to the centre of knee joints, evaluates and analyses the results obtained, and points out the drawbacks in this approach. Section 4.3 describes the approach for localising the ROI directly, and shows the results obtained. Section 4.4 presents a comparison against the baseline approaches discussed in Chapter 3 and the proposed methods in this chapter for automatically localising the knee joints in the X-ray images. Section 4.5 summarises the work in this chapter.

4.2 Localisation with Reference to Knee Joint Centre

In the initial approach to localise the knee joints in X-ray images using a FCN, a similar strategy to template matching and the SVM based methods is followed; that is to detect the centre of knee joints and to extract the ROI with reference to the detected centres. Figure 4.1 shows the steps involved in this method: training a FCN to detect the knee joint centres (20×20 pixels), computing the coordinates of

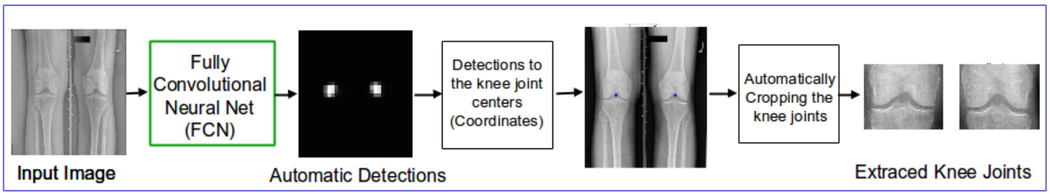

Input Image
Fully Convolutional Neural Net (FCN)
Automatic Detections
Detections to the knee joint centers (Coordinates)
Automatically Cropping the knee joints
Extracted Knee Joints

Figure 4.1: Automatic localisation of knee joints with reference to the centre of the knee joints.

the centres from the FCN output, and extracting a fixed size region as knee joints. In the next section, the experimental data and the ground truth used to train the FCNs are introduced.

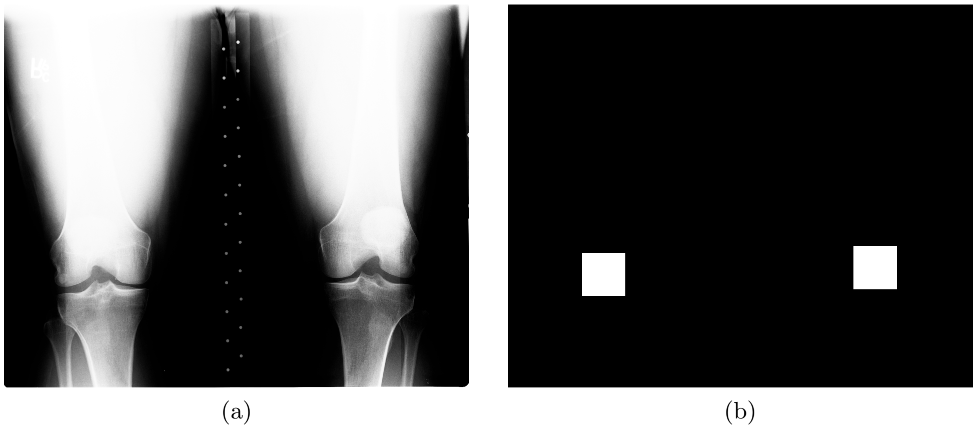
4.2.1 Dataset and Ground Truth Generation

The data used for the experiments are taken from the baseline cohort of the OAI dataset. In total 4,446 X-ray images are selected from the entire dataset based on the availability of KL grades for both knee joints. The knee joint centres in all these X-ray images are manually annotated, after downscaling to 10% of the actual size. Binary masks of size 20×20 pixels are marked around the knee joint centres using the annotations. Figure 4.2 shows an instance of an input X-ray image and the binary mask annotations corresponding to the knee joint centres. The image patches from the masked region i.e. the knee joint centres, are taken as positive training samples and the patches from rest of the image are taken as the negative training samples to train an FCN. The dataset is split into training (3,333 images) and test (1,113 images) sets.

4.2.2 Training Fully Convolutional Neural Networks

Initial Configuration and Training

To start, a FCN is configured with a lightweight architecture containing 4 convolutional layers followed by a fully convolutional layer, which is a convolutional layer with a kernel size [1×1] and that uses a *sigmoid activation*.



(a)

(b)

Figure 4.2: (a) An input X-ray image and (b) The binary mask annotations for knee joint centres.

Table 4.1: Initial FCN Configuration for detecting the knee joint centres.

Layer	Kernel	Kernel Size
Conv1	32	3×3
Conv2	32	3×3
Conv3	64	3×3
Conv4	64	3×3
Conv5	1	1×1

FCNs use fully convolutional layers at the last stage to make pixel-wise predictions [88]. Table 4.1 shows the network configuration in detail. Each convolution layer is followed by a ReLU layer.

The network parameters are trained from scratch with training samples of knee OA radiographs from the OAI dataset. The dataset is split into training (3,333 images) and test (1,113 images) sets. The ground truth for training the network are binary images with masks specifying the ROI: the knee joints. The network is trained to minimise the total *binary cross entropy* between the predicted pixels and the ground truth. *Stochastic gradient descent* (SGD) with default parameters: learning rate = 0.01, decay = $1\mathrm{e}^{-6}$, momentum = 0.9, and nesterov = True, is used. The network is trained for 40 epochs and the batch size is 10. Figure 4.3 shows an instance of the test input, the ground truth and the output (pixel-wise predictions) of the FCN. From the predictions of this FCN, it is observed that the network is able

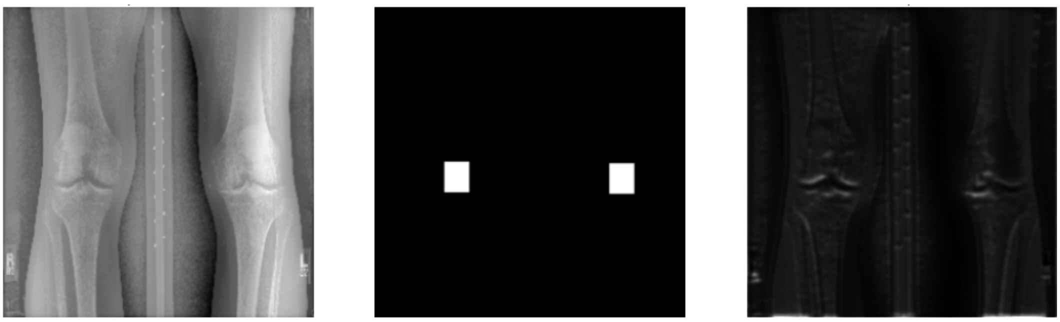


Figure 4.3: An instance of input, ground truth and output (predictions) of FCN.

to slightly detect the edges of the knee joints and these are promising initial results. In an attempt to improve the detections, the FCN configurations are experimented and for this the hyper-parameters of the network are tuned.

Receptive Field

When dealing with high-dimensional inputs such as images, it is impractical to connect neurons in the current level to all the neurons in the previous volume. Instead, each neuron is only connected to a local region of the input volume. The spatial extent of this connectivity is a hyper-parameter called the receptive field of the neuron [59]. The receptive field size, otherwise termed the effective aperture size of a CNN, shows how much a convolutional node sees of the input pixels (patch) that affects a node's output. The effective aperture size depends on kernel size and strides of the previous layers. For instance, a 3×3 kernel can see a 3×3 patch of the previous layer and a stride of 2 doubles what all succeeding layers can see.

The receptive field size of neurons in the final layer of the FCNs is calculated and used to analyse the output of FCNs and the overall detection results. The receptive field size of a neuron in the final layer (Conv5) of the initial FCN configuration (Table 4.1) is 9, which is low and may be a reason for poor performance of this network. Larger convolutional kernel sizes to increase the receptive field of the network is investigated. Section 4.2.5 shows that a network (Table 4.6) with larger receptive field gives the best results for detecting the knee joint centres.

Table 4.2: FCN for detecting the knee joint centres.

Layer	Kernel	Kernel Size
Conv1	32	7×7
Conv2	64	3×3
Conv3	96	3×3
Conv4 (fullyConv)	1	1×1

Tuning the FCN Hyper-parameters

VGG-M-128 [63], the deep convolutional neural network developed by the Oxford visual geometry group (VGG) uses kernel size 7×7 in the first convolutional layer and 5×5 in the following convolutional layer. Inspired by this, kernel sizes of 5×5, and 7×7 for the first convolutional layer are tested retaining the other settings. The kernel size 7×7 gives better results in this configuration. This is because of the larger receptive field size of the 7×7 kernel in comparison to the 3×3 kernel.

Next, the experiments are conducted by varying the number of convolutional layers and also the number of filters (kernel) in a convolutional layer, before obtaining the configuration that gave the best results based on visual observations. Table 4.2 shows the configuration of the network derived from the initial configuration and the receptive field size of a neuron in the final layer (Conv4) is 11. The networks are trained with 3,333 images and tested on 1,113 images from the OAI dataset.

There is an improvement in the detections using this network in comparison to the previously tested configurations. Figure 4.4 shows an instance of the output predictions of this network. To quantitatively evaluate the automatic detections, the well-known Jaccard Index is used.

4.2.3 Quantitative Evaluation

A simple contour detection is used and the Jaccard index i.e. the overlap statistics calculated by the Intersection over Union (IoU) to evaluate the automatic detections of the FCN. The steps involved are as follows:

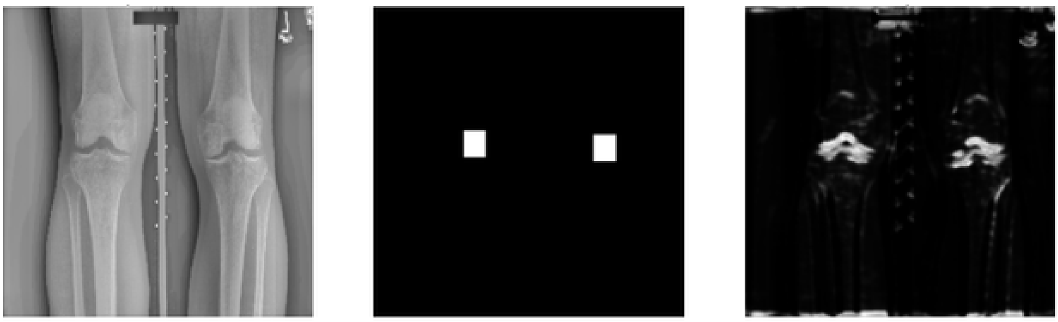


Figure 4.4: An input image, ground truth, and outcome of the final FCN.

- First, the objects are detected i.e. the knee joint regions from the output image of the FCN using simple contour detection [92]. Contours can be explained simply as a curve joining all the continuous points (along the boundary), having the same colour or intensity. The contours are a useful tool for shape analysis and simple object detection and recognition. In this method, first the images are converted to binary by applying Otsu's threshold. Next, the contours of the objects or shapes in the binary image are automatically detected and recorded [92].
- Next, the detected objects in the image are sorted based on the area and from these the top two are selected. This is to eliminate noise or other faint edges picked up by the FCN.
- The centroids of the largest two detected regions are recorded as the knee joint centres.
- A binary mask of 20×20 pixels size is marked around each detected knee joint centre.
- The Jaccard index is computed for each image with the masks of predicted centres and the masks predefined using manual annotation i.e. the labels used for training FCN.

In total 1,113 X-ray images that is 2,226 knee joints are included in the test set. The FCN with the final configuration detects 1,851 knee joints in the test set

Table 4.3: FCN with pooling and up-sampling layers.

Layer	Kernel	Kernel Size	Strides
Conv1	32	7 ×7	1
MaxPool2	–	2×2	2
Conv3	64	3×3	1
MaxPool4	–	2×2	2
Conv5	96	3×3	1
UpSamp6	–	4×4	1
Conv7 (fullyConv)	1	1×1	1

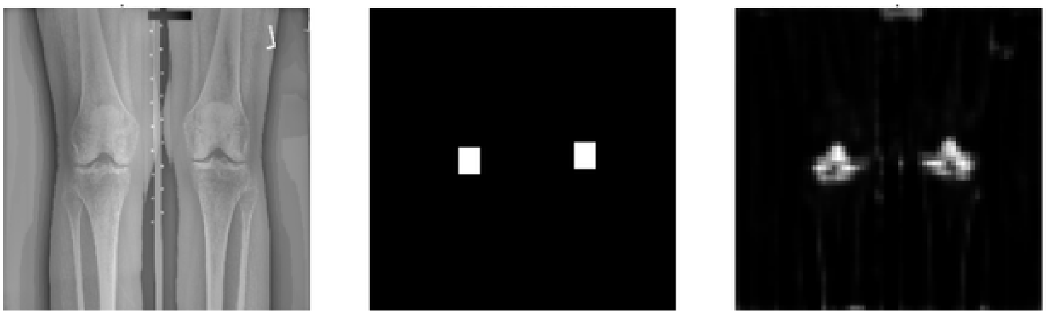


Figure 4.5: Prediction of the FCN with max pooling and up-sampling layers.

with Jaccard index ≥ 0.5, the accuracy of detection is **83.2%** with a mean 0.66 and standard deviation 0.18. This is an improvement in comparison to previous approaches but still falls short of perfect detections. The pooling and up-sampling layers in the FCN are varied and experimented in an attempt to improve the detection accuracy. This will help to increase the receptive field size and in turn improve the overall detections.

4.2.4 FCN with Pooling and Up-sampling Layers

Two max pooling layers with stride 2 and up-sampling by a factor of 4 are included to the previous configuration (Table 4.2). Table 4.3 shows the FCN architecture in detail. Each convolutional layer is followed by a ReLU activation.

Figure 4.5 shows the output of this network for a test image. On visual observation, the output image contains less noise and the detections are improving compared to the previous approaches, even though the output image resolution is

Table 4.4: FCN with 3 Convolution-Pooling stages for detecting the knee joint centres.

Layer	Kernel	Kernel Size	Strides
Conv1	32	7×7	1
MaxPool2	–	2×2	2
Conv3	32	3×3	1
MaxPool4	–	2×2	2
Conv5	64	3×3	1
MaxPool6	–	2×2	2
Conv7	96	3×3	1
UpSamp8	–	8×8	1
Conv9 (fullyConv)	–	1×1	1

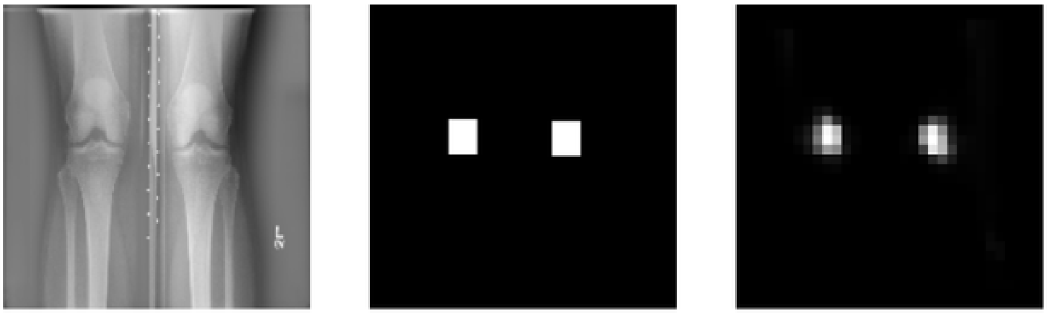


Figure 4.6: Predictions of the FCN with 3 Convolution-Pooling stages.

low. This is due to the inclusion of pooling and up-sampling stages to the network and this has increased the receptive field size of the final layer (Conv7) to 34. The number of convolutional-pooling stages is increased, to see if there is improvement in the detections. Table 4.4 shows the architecture of this network in detail.

From the output of this FCN, it can be observed that the detections become more precise in comparison to the previous networks even though the resolution is low in comparison to the previous networks. Figure 4.6 shows an instance of the input test image, ground truth and the FCN output.

The outcomes of this FCN are evaluated using Jaccard index and the detection accuracy is **96.7%**, that is in total 2,152 out of 2,226 knee joints are detected with a Jaccard index ≥ 0.5. The Jaccard index mean is 0.74 and standard deviation is 0.13. The detection accuracy is high in comparison to the previous networks. Table 4.5 shows the detection accuracy of the FCN for the Jaccard index values at 0.25, 0.5 and 0.75.

Table 4.5: Detection accuracy of FCN based on Jaccard Index.

Jaccard Index	JI $\geq$ 0.25	JI $\geq$ 0.5	JI $\geq$ 0.75
Detection Accuracy	98.5 %	96.7 %	39.6 %

This FCN (Table 4.4) has three convolutional-pooling stages. A configuration with 4 convolutional-pooling stages followed was tested by adding an up-sampling layer with kernel size (16×16). There was no improvement in the detection accuracy for this configuration.

4.2.5 Best Performing FCN for Detecting the Knee Joint Centres

Before settling on the final architecture, experiments were done by varying the number of convolution stages, the number of filters and kernel sizes in each convolution layer. The best performing FCN (Table 4.6) was selected based on a high detection accuracy on the test data. This network was trained with the OAI dataset containing 4,444 knee radiographs. The dataset was split into a training set containing 3,333 knee images and test set containing 1,113 knee images. The validation set (10%) was taken from the training set. The effective aperture size of this FCN (Table 4.6) for a node in the last convolutional layer (before up-sampling) is 66. The aperture size for the previous networks shown in Table 4.4 is 42 and Table 4.3 is 34. For the other tested configurations the effective aperture size is even lower (less than 30).

Table 4.6 shows the configuration of the best performing FCN for detecting the knee joint centres. This FCN is based on a lightweight architecture and the network parameters (in total 214,177) are trained from scratch. The network consists of 4 stages of convolutions with a max-pooling layer after each convolutional stage, and the final stage of convolutions is followed by an up-sampling and a fully-convolutional layer. The network uses a uniform [3×3] convolution and [2×2] max pooling. Each convolution layer is followed by a ReLU

Table 4.6: Best performing FCN for detecting the knee joint centres.

Layer	Kernel	Kernel Size	Strides
Conv1	32	3×3	1
MaxPool1	–	2×2	2
Conv2_1	32	3×3	1
Conv2_2	32	3×3	1
MaxPool2	–	2×2	2
Conv3_1	64	3×3	1
Conv3_2	64	3×3	1
MaxPool3	–	2×2	2
Conv4_1	96	3×3	1
Conv4_1	96	3×3	1
UpSamp5	–	8×8	1
Conv5 (fullyConv)	–	1×1	1

activation layer. After the final convolution layer, an [8×8] up-sampling is performed as the network uses 3 stages of [2×2] max pooling. The up-sampling is essential for an end-to-end learning by back propagation from the pixel-wise loss and to obtain pixel-dense outputs [88], when pooling layer(s) and strides more than one are used in the network. The final layer is a fully convolutional layer with a kernel size of [1×1] and uses a sigmoid activation for pixel-based classification. The input to the network is of size [256×256].

This network was trained to minimise the total binary cross entropy between the predicted pixels and the ground truth using *stochastic gradient descent* (SGD) with default parameters: learning rate = 0.01, decay = $1e^{-6}$, momentum = 0.9, and nesterov = True. This network was trained for 40 epochs with a batch size 32. The validation (10%) data was taken from the training set. Figure 4.7 shows the learning curves when training this network and decrease in the validation and training losses.

Table 4.7 shows the results of the best performing FCN. This network achieved a detection accuracy of **97.1%**, in total 2,162 knee joints out of the 2,226 test samples detected with a Jaccard index 0.5. The Jaccard index mean is 0.76 and standard deviation is 0.12.

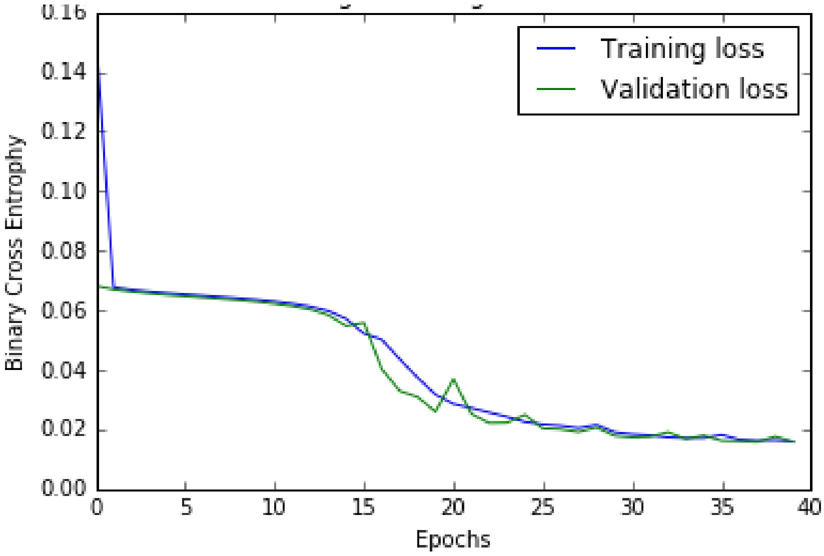


Figure 4.7: Training and validation losses of the FCN.

Table 4.7: Detection accuracy of the best performing FCN.

Jaccard Index	JI $\geq$ 0.25	JI $\geq$ 0.5	JI $\geq$ 0.75
Detection Accuracy	98.9 %	97.1 %	43.3 %

4.2.6 Error Analysis

The results of the best performing FCN (Table 4.7) show 99% detection accuracy for a Jaccard index $\geq$ 0.1, in total 2,205 out of 2,226 knee joints are successfully detected. On observing the failed detections: 1% (in total 21 knee joints), there are two patterns.

1. The output of the FCN is very faint or no detections at all. Figure 4.8 shows two instances of input X-ray images, masks defining the knee joint centres as ground truth, and output of the best performing FCN with faint detections. The input images with variations in the local contrast and local luminance due to the imaging protocol variations are the main cause for this error. Histogram equalisation is used as a pre-processing step to adjust the contrast of the input images. Even though this adjusts the contrast globally in an image, there are still contrast variations in portions of the image. Local contrast

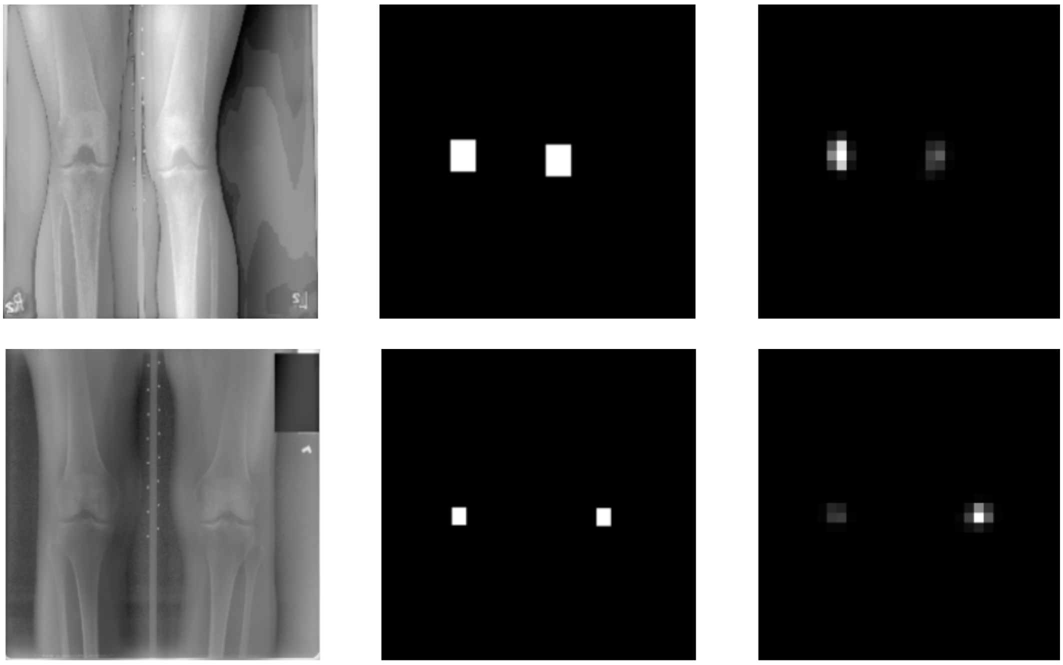


Figure 4.8: Error analysis: X-ray images, ground truth, FCN output - weak detections

enhancement algorithms [93] or adaptive histogram equalisation [94] can be used to normalise the images for variations in the local contrast and local luminance.

2. The FCN output picks up noise along with the knee joints. Figure 4.9 shows two instances of input X-ray images, masks defining the knee joint centres as ground truth, and output of the best performing FCN with noise. The reason for this error is due to the variations in the imaging protocol and resolution of the X-ray images, and presence of artefacts in the input X-ray images. Intuitively, the FCN uses horizontal edge detection along with other features to detect the knee joints. The artefacts with predominant horizontal edges are picked up by the FCN along with the centre of knee joints. When simple contour detection is applied on the FCN output, instead of the knee joints the artefacts are also detected.

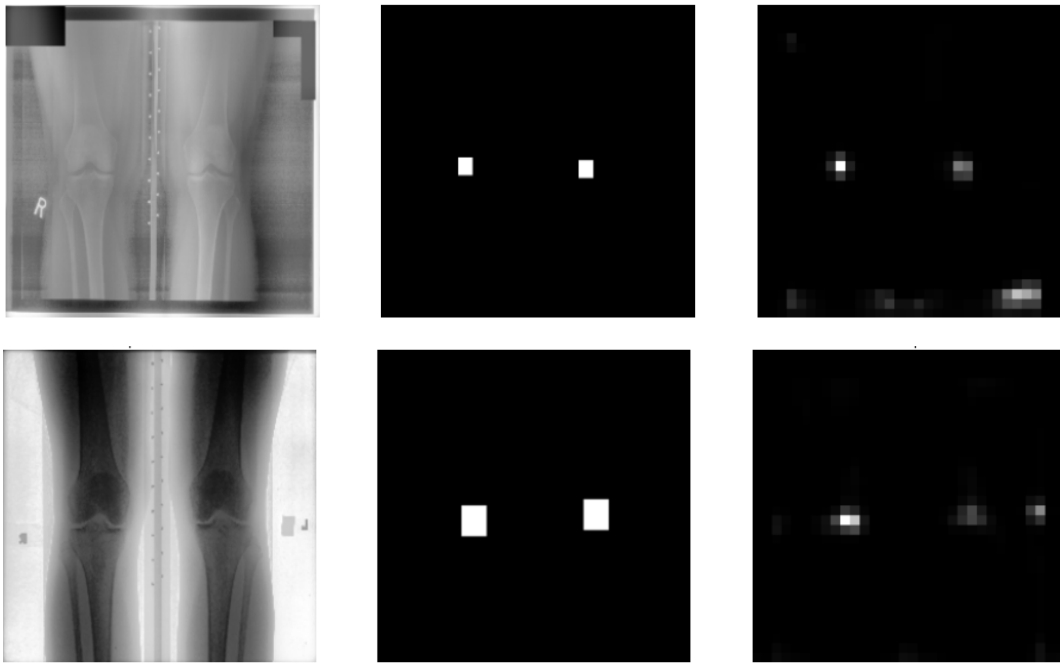


Figure 4.9: Error analysis: X-ray images, ground truth, FCN output - detections with noise.

4.2.7 Automatically Extracting the Knee Joints

After training FCNs to automatically detect the centre of the knee joints, the next step is to extract the ROI i.e. the knee joints with reference to the detected centres. The initial goal is to train an end-to-end network for localising the knee joints i.e. to directly predict the bounding box co-ordinates of the knee joints from the input X-ray images. A bounding box regression is investigated [95] that is a network trained on top of the FCN (Table 4.6) output, to achieve this. First, CNNs are trained with the masks (20×20) of knee joint centres as the input (256×256) and the bounding box coordinates of the left knee joint (x_1, y_1) and right knee joint (x_2, y_2) as the ground truth (labels). Next, CNNs are trained with the X-ray images as input and the targets (labels) are the bounding box coordinates instead of the binary masks. However in both the experiments, the networks trained to predict the bounding boxes give low accuracy. On considering the overall knee joint centres, there is no large variations in the centre coordinates. The reason for the low accuracy is that the networks are not learning discernible features to predict the bounding box coordinates. This affects the overall performance of the localisation. Therefore, a simple approach based on contour detection is used to calculate the centres and extract the knee joints. Figure 4.10 shows an X-ray image with the centres, the left and the right knee joints extracted from the X-ray image using the centroids. The steps involved in this method are as follows.

- First, the contour detection [92] is used on the FCN output to calculate the spatial coordinates of the knee joint centres. In the contour detection method, first the input images (FCN output) are converted to binary by applying Otsu's threshold. Next, the contours from the binary image are automatically detected and recorded. Finally, the centroids are calculated from the detected knee joint regions.
- The knee OA radiographs are resized to 2560×2560, that is 10 times the size of the FCN output 256×256.
- The detected knee joint centres are up-scaled to a factor of 10.
- Fixed size regions (640×560) are extracted around the up-scaled centres as the knee joint regions. After testing and visualising different sizes for the knee joint crop, image patch with the size (640×560) is found to be mostly suitable and containing the required ROI for further quantification. Figure 4.10 shows an instance of the extracted left and right knee joints.

4.2.8 Localisation Results

The results of the FCN are compared to the previous methods: template matching and SVM-based method to automatically detect the centre of the knee joints. All these methods are evaluated based on the Jaccard index (JI). Table 4.8 shows the detection accuracy of the knee joint centres using FCN, SVM-based method, and template matching. The results show that the proposed method using FCN clearly

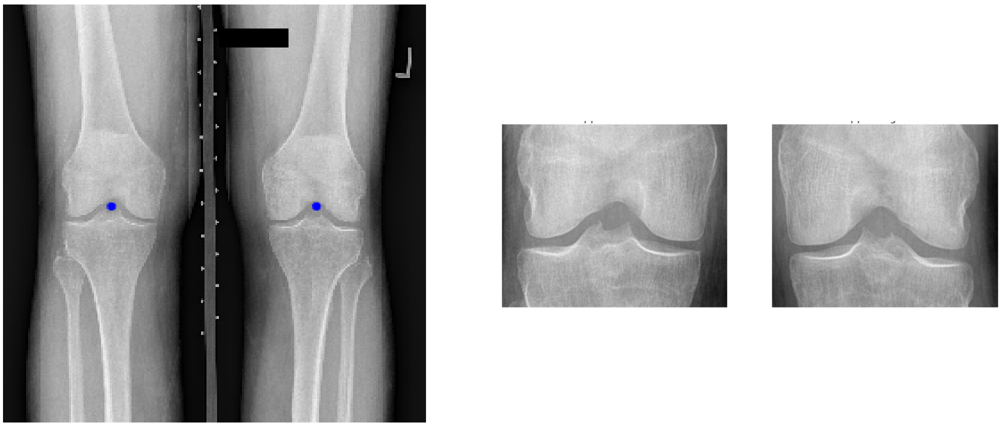


Figure 4.10: A knee X-ray image with the detected centres and the extracted left and right knees.

Table 4.8: Comparison of methods used for localising the centre of the knee joints

Method	JI > 0	JI ≥ 0.5	JI ≥ 0.75	Mean	Std. Dev.
Template Matching	54.4%	8.3%	3.1%	0.1	0.2
SVM-based Method	81.8%	38.6%	10.2%	0.36	0.31
Fully ConvNet	**98.9%**	**97.1%**	**43.3%**	**0.76**	**0.12**

outperforms the previous methods. This also demonstrates that feature learning using an FCN is a better approach for detecting the knee joints than using hand-crafted features such as Sobel gradients and the template matching method that is sensitive to intensity level variations. However, the extracted knee joints from this method have some limitations.

4.2.9 Limitations of this Method

In all three approaches; FCN-based, SVM-based and template matching, the centre of the knee joints are detected and these are used as reference for automatically localising the knee joints. There are some limitations in extracting a fixed size region as the ROI with reference to the detected centres due to the variations in the resolution of the X-ray images and the variations in the size of the knee joints.

All the images are resized to a fixed size 2,560×2,560 and extract a fixed size region 640×560 around the detected centres as the ROI. Due to this scaling issue,

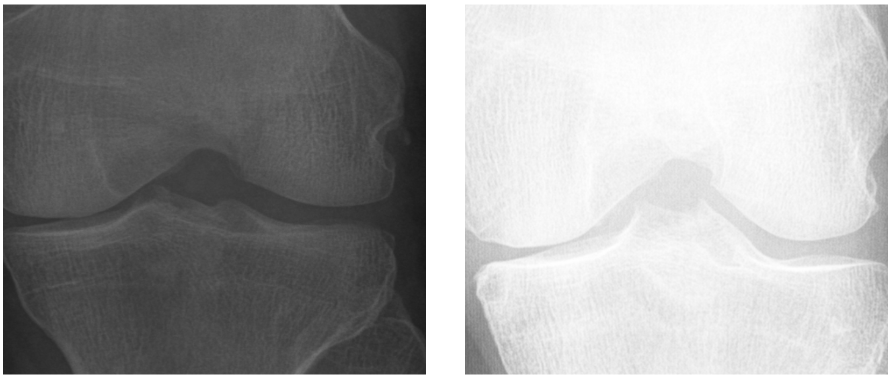


Figure 4.11: Anomalies in the automatic extraction of the ROI.

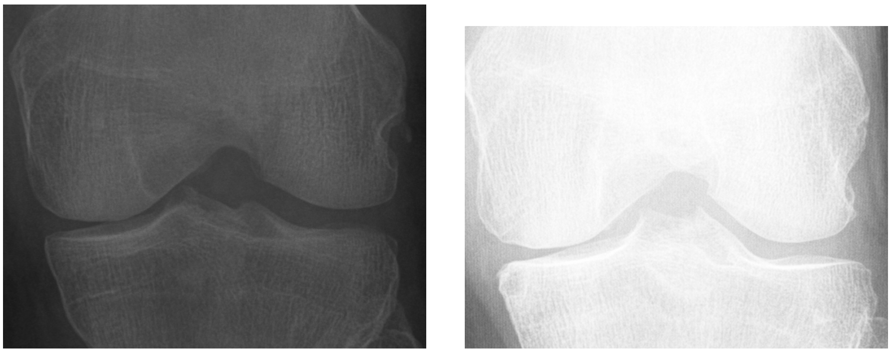


Figure 4.12: The actual ROI for the knee joints in Figure 4.11.

portions of the knee joints are omitted in the automatic extraction of the ROI. Figure 4.11 shows such instances. Figure 4.12 shows the corresponding actual ROIs.

Due to the varying sizes of the knee joints and a fixed size region being extracted as the ROI, there are differences in the aspect ratio of the extracted and the actual ROI. Figure 4.13 shows instances where the knee joints are small in comparison to the fixed size region extracted as the ROI. Figure 4.14 shows the actual ROIs.

The classification of the automatically extracted knee joints is compared to the manually extracted knee joints. There is a decrease in the accuracy by a margin of 3–4% when using the automatically extracted knee joints with reference to the detected centres. The discrepancies in the localisation of knee joints affects the overall classification of the knee OA images. To overcome these limitations, as the next approach FCNs are trained to detect the ROI itself, instead of detecting the knee joint centres.

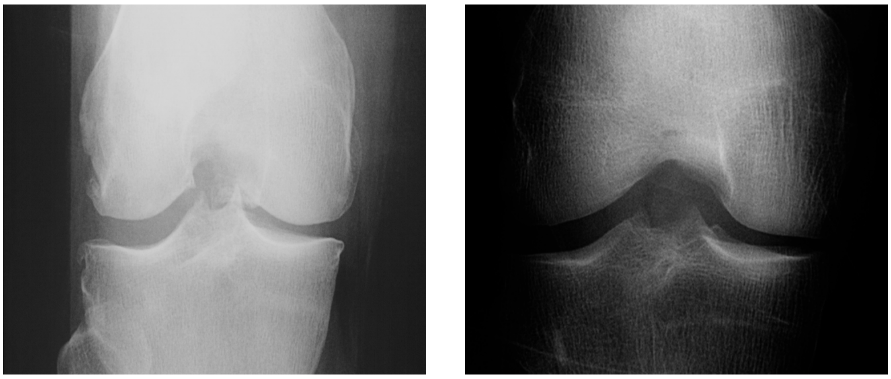


Figure 4.13: Variations in the aspect ratio of the extracted knee joints.

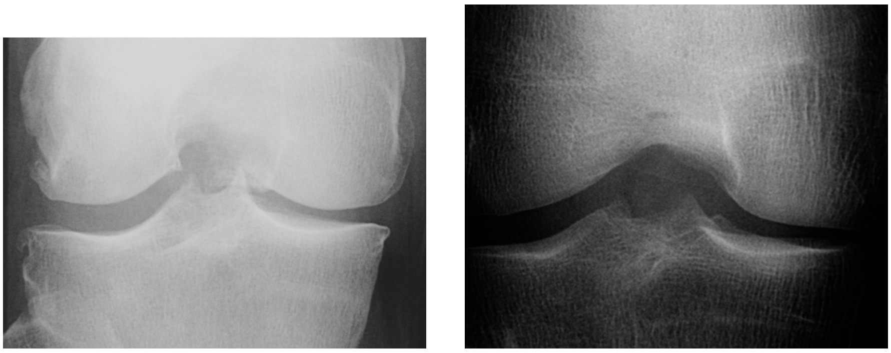


Figure 4.14: The actual ROI for the extracted knee joints in Figure 4.13.

4.3 Localising the Region of Interest

The previous methods to localise the knee joints in the X-ray images with reference to the automatically detected centres have certain limitations. To overcome these limitations and to improve the localisation, FCNs are trained to detect the ROI directly. Figure 4.15 shows the steps involved in this method.

4.3.1 Dataset and Ground Truth

For the experiments in this approach, a new dataset from the MOST is used along with the data from the previous experiments, the baseline cohort of the OAI dataset. In total 4,446 X-ray images are selected from the OAI dataset and 2,920 X-ray images from the MOST dataset based on the availability of KL grades for both

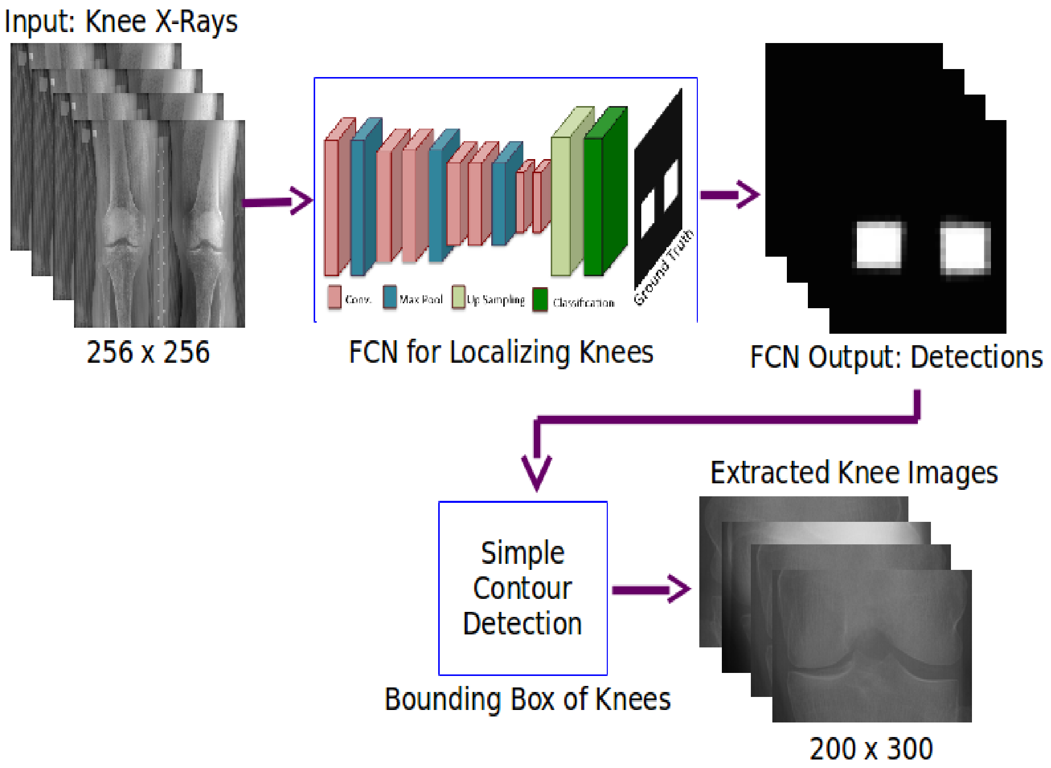
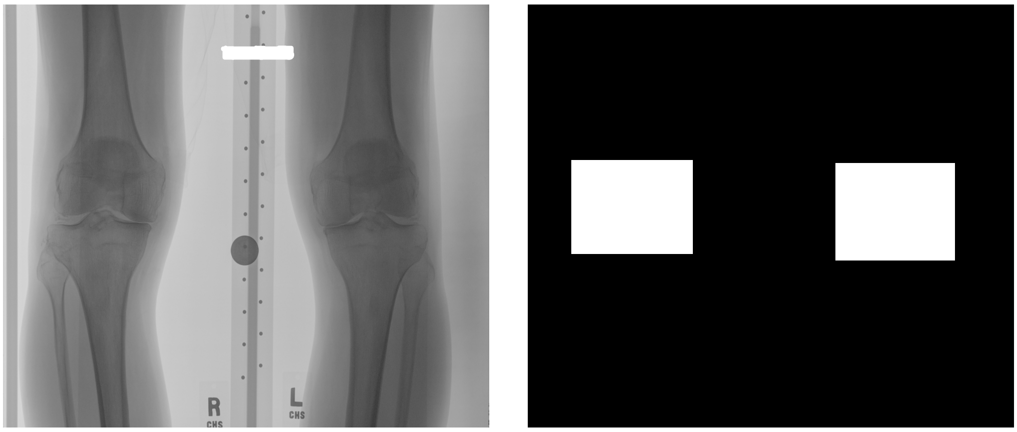


Figure 4.15: Automatic localisation of the Region of Interest.

knee joints. The full ROI is manually annotated in all these X-ray images, after downscaling to 10 % of the actual size. The down-sampling of the images is necessary to reduce the computational costs. Binary masks are generated based on the manual annotations. Figure 4.16 shows an instance of an input X-ray image and the binary mask annotations corresponding to the ROI. The image patches from the masked region i.e. the knee joints are taken as positive training samples and the patches from rest of the image are taken as the negative training samples to train a FCN. The datasets are split into a training/validation set (70%) and test set (30%). The training and test samples from the OAI dataset are 3,146 images and 1,300 images, and from the MOST dataset are 2,020 images and 900 images.



(a) (b)

Figure 4.16: (a) An input X-ray image and (b) The binary mask annotations for the region of interest.

4.3.2 Training the FCN

First, a FCN is trained using the same architecture (Table 4.6) from the previous approach to detect the ROI. Initially, the network is trained with training samples from OAI dataset and test it with OAI and MOST datasets separately. Next, the training samples are increased by including the MOST training set where the test set is a combination of both OAI and MOST test sets. This network is trained to minimise the total binary cross entropy between the predicted pixels and the ground truth using the adaptive moment estimation (Adam) optimiser [96] with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\mathrm{e}^{-8}$. Adam optimiser gives faster convergence than standard SGD. Figure 4.17 shows the learning curves converging to small loss when training this network. Figure 4.18 shows the output of this network for a test image.

A few other network configurations are tested by varying the number of convolutional-pooling stages, convolutional layers in each stage and the number of convolutional kernels in a convolutional layer. There was no further improvement in the detection accuracy on the validation set. Therefore, this configuration was settled as the final network for localising the knee joints.

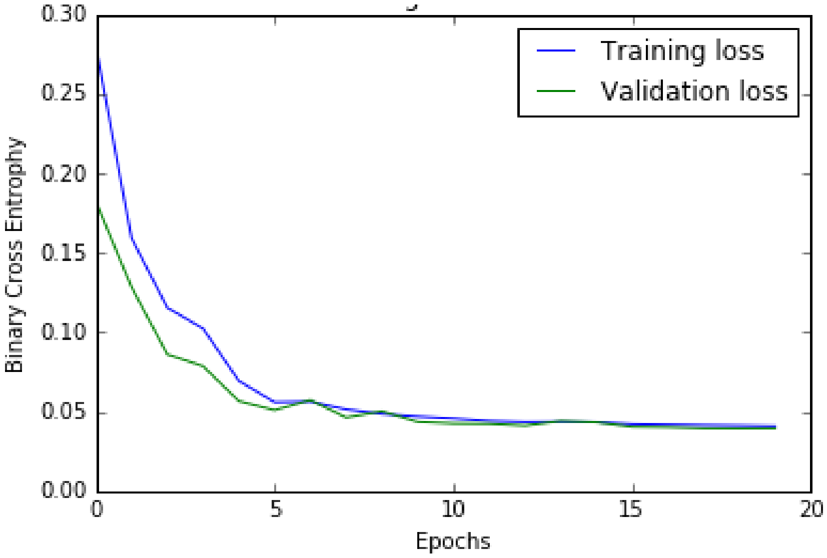


Figure 4.17: Training and validation losses of the FCN.

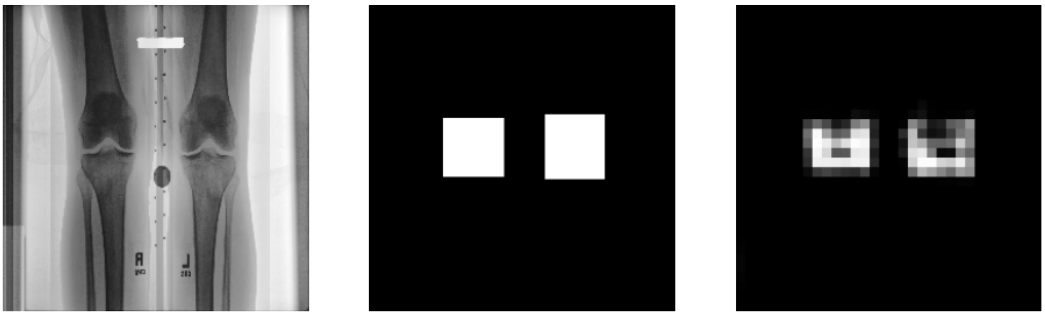


Figure 4.18: An input X-ray image, ground truth and output prediction of the FCN.

4.3.3 Quantitative Evaluation

The Jaccard index, i.e. the intersection over Union (IoU) of the automatically detected and the annotated knee joint is used to quantitatively evaluate the automatic detections. For this evaluation, all the knee joints in both the OAI and MOST datasets are manually annotated using a fast annotation tool. Table 4.9 shows the number (percentage) of knee joint correctly detected based on the Jaccard index (JI) values greater than 0.25, 0.5 and 0.75 along with the mean and the standard deviation of JI. Table 4.9 also shows detection rates on the OAI and MOST test sets separately.

Considering the anatomical variations of the knee joints and the imaging

Table 4.9: Comparison of automatic detection based on the Jaccard Index (JI).

Test Data	JI $>$ 0	JI $\geq$ 0.5	JI $\geq$ 0.75	Mean	Std. Dev.
OAI	100%	100%	88%	0.82	0.06
MOST	99.7%	98.8%	80.6%	0.80	0.09
Combined OAI-MOST	**100%**	**100%**	**92.2%**	**0.83**	**0.06**

protocol variations, the automatic detection with a FCN is highly accurate with 100% detection accuracy for JI≥0.5 and 92.2% (4,056 out of 4,400) of the knee joints for J≥0.75 being correctly detected. Further evidence is provided to show that the FCN based detection is highly accurate by showing that the quantification results obtained with the automatically extracted knee joints giving results on par with manually segmented knee joints in the next chapter, Section 5.3.8.

4.3.4 Qualitative Evaluation

Figures 4.19, 4.20, and 4.21 show a few instances of successful knee joint detections with the JI values for the left and right knee detections. Detecting the ROI directly gives high accuracy (100%) in comparison to the previous method (Section 4.2) to detect the knee joint centres and extracting a fixed size region as the ROI. The FCN in this method learns features from a relatively larger region (the actual ROI) in comparison to the previous method where the FCN is confined to learn features from a small region (20×20), the centre of the knee joints, and therefore, the detections are more accurate.

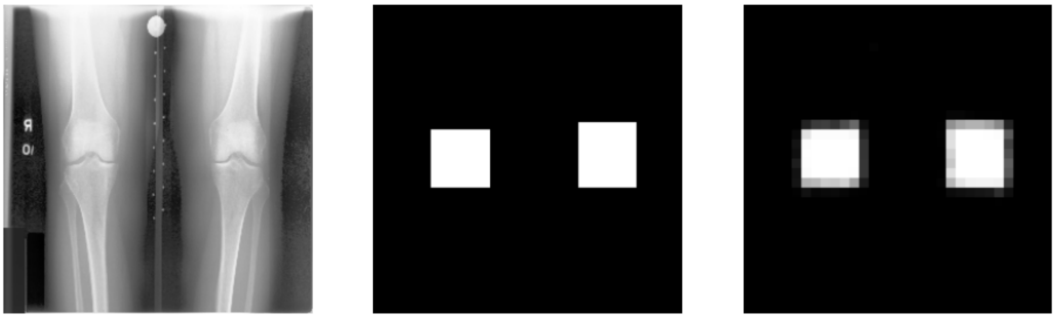


Figure 4.19: Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.98, right knee with JI=0.888.

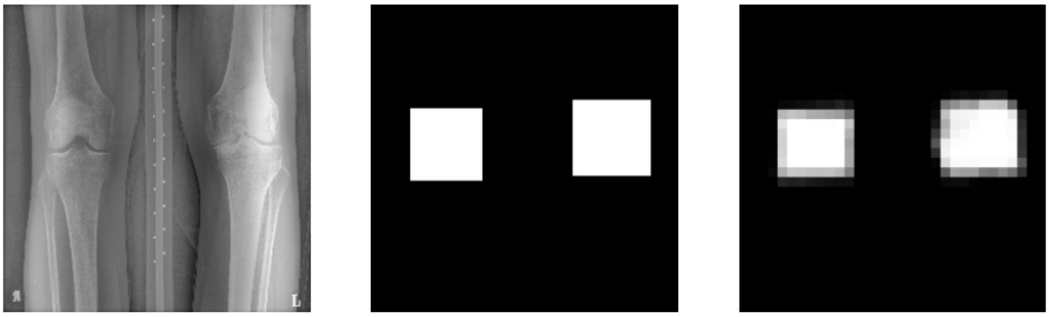


Figure 4.20: Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.879, right knee with JI=0.969.

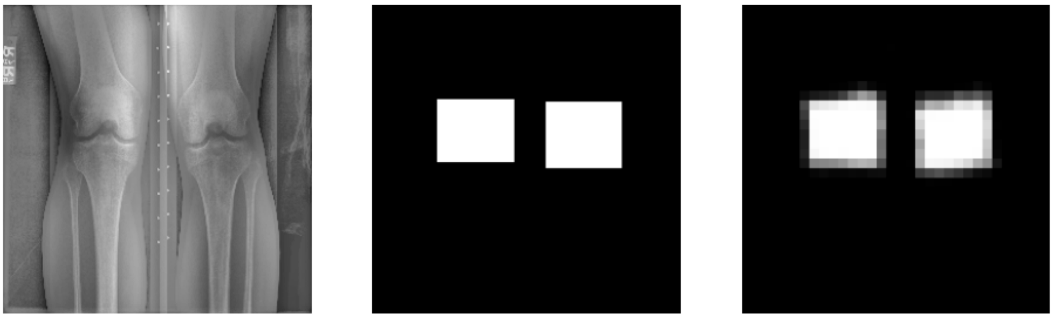


Figure 4.21: Qualitative Evaluation: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.768, right knee with JI=0.984.

4.3.5 Error Analysis

This method is highly accurate with 100% detection accuracy for a JI $\geq$ 0.5. Nevertheless, there are a few anomalies in the FCN detections due to variations in the imaging protocols, presence of artefacts and noise in the input images. Figures 4.22 and 4.23 show two instances where one knee has undergone joint-arthoplasty and the knee implants are visible in the X-ray images, and due to this the FCN detections are distorted. Figures 4.24, 4.25 and 4.26 show a few instances of X-ray images with noise and presence of artefacts due to imaging protocols. This adversely affects the FCN detections.

4.3.6 Extracting the Knee Joints

The bounding boxes of the knee joints are calculated using simple contour detection from the output predictions of the FCN. After converting the FCN output to binary

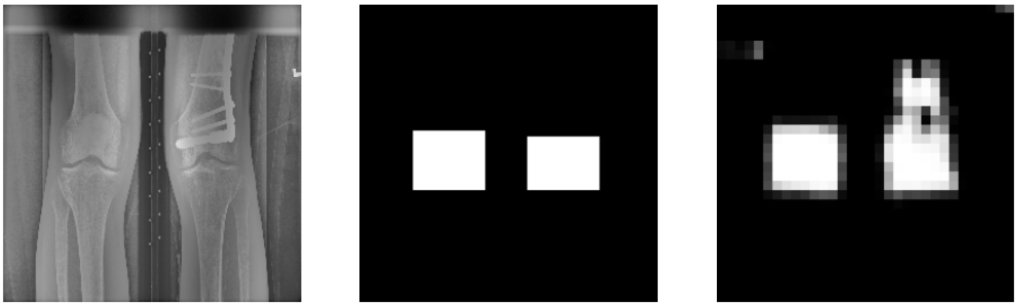


Figure 4.22: Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.83, right knee with JI=0.398. The implants in the right knee is the reason for this localisation error.

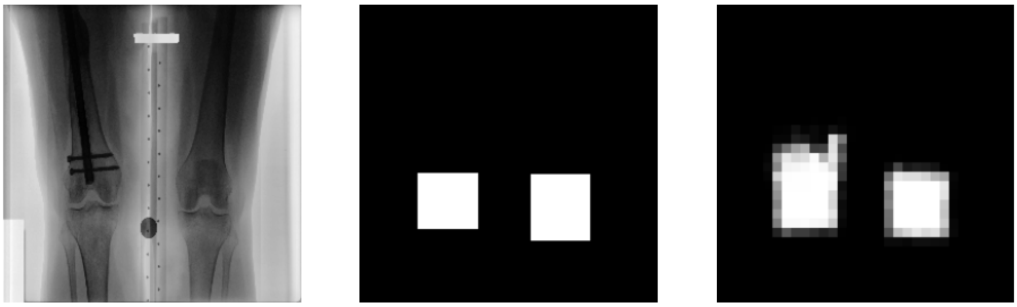


Figure 4.23: Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.473, right knee with JI=0.837. The implants in the left knee is the reason for this localisation error.

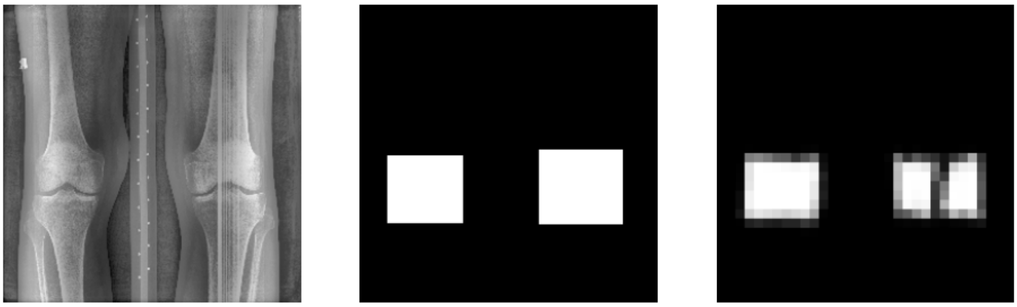


Figure 4.24: Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.887, right knee with JI=0.356. The noise in the right knee causes this localisation error.

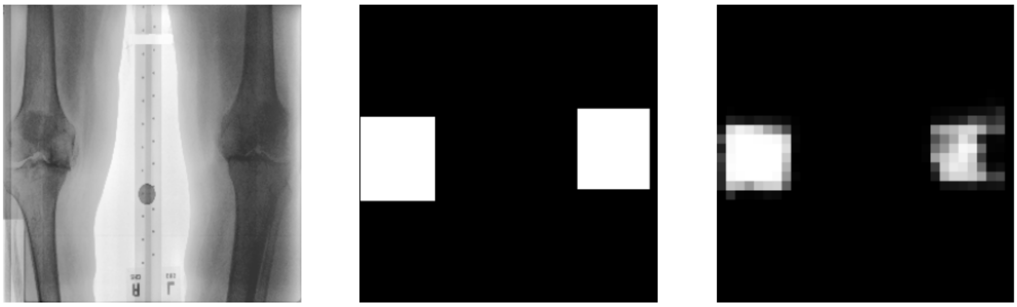


Figure 4.25: Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.681, right knee with JI=0.488. The localisation error in this image is due to the variation in the imaging protocol.

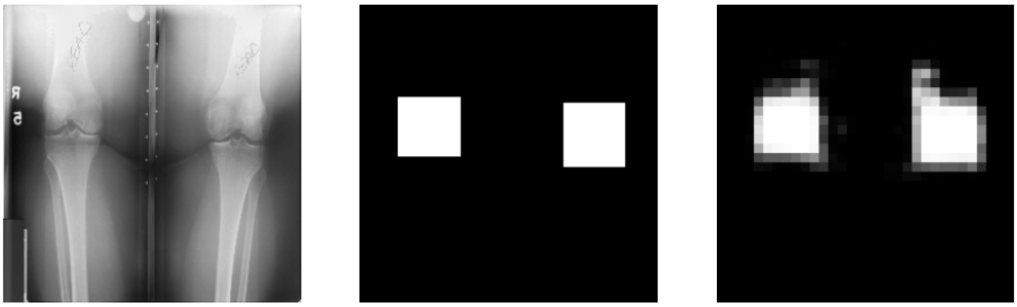


Figure 4.26: Error Analysis: An input X-ray image, ground truth, and FCN detections: left knee with JI=0.768, right knee with JI=0.507. The variations in the local contrast and luminance affects the localisations.

image using Otsu's threshold, the contours are detected using simple image analysis by calculating the zero order moments [92], which gives the perimeter of the detected object. The contours are recorded as bounding boxes. The knee joints are extracted from knee OA radiographs using the bounding boxes. The bounding boxes are upscaled from the output of the FCN that is of size [256 × 256] to the original size of each knee OA radiograph, before extracting the knee joints so that the aspect ratio of the knee joints is preserved.

4.4 Conclusion

Automatically localising the knee joints in X-ray images is an important and an essential step before quantifying knee OA severity. Previously, template matching was implemented as a baseline method to localise the knee joints, proposed by Shamir et al. [2, 7], and it was shown that the detection accuracy is low ($\sim 30\%$) in this method for large datasets like OAI. To improve the localisation, a SVM-based method with Sobel horizontal image gradients as features is proposed in this thesis. This method showed a large improvement in detection accuracy (82%) but still falls short of perfect localisation. The anomalies in localised knee joints can affect the following step: classifying the localised knee joints to quantify knee OA severity.

Instead of using hand-crafted features, a deep learning-based solution is proposed in this chapter to further improve localisation. FCNs were trained to automatically detect and extract the knee joints. All three methods: template matching, SVM-based and FCN-based are evaluated using a common metric: the Jaccard Index. This method achieves almost perfect detection with 100% accuracy for a Jaccard Index 0.5 and an accuracy of 92% for a Jaccard index greater than equal to 0.75. The author believes this performance is sufficient to localise and extract the knee images for classification. As such the further improvements are left as future work. The localisation performance may be improved by including additional pre-processing steps to remove the artefacts and noise in the images, and to normalise the local contrast variations in the images. Using additional data for learning and data augmentation may improve the localisation performance.

4.5 Chapter Summary

In this chapter, two approaches for automatically detecting and localising the knee joints in X-ray images using FCNs are introduced. As a first approach, FCNs are trained to automatically detect the knee joint centres and a fixed size region is cropped as the ROI with reference to the detected centres. Though this approach achieved high detection accuracy, the extracted knee joints had certain limitations due to the variations in the resolution of the knee radiographs and the anatomical variations of the knees. To overcome these drawbacks and to further improve the localisation, as the second approach FCNs are trained to directly localise the ROI instead of knee joint centres. The results from this method are near perfect and outperform the previous methods to localise the knee joints. The next chapter introduces the proposed methods for quantifying knee OA severity through classification and regression on the localised knee joints.

Chapter 5

Automatic Quantification of Knee OA Severity

5.1 Introduction

The previous chapter focused on developing and evaluating a deep learning based framework to automatically localise the knee joints from the X-ray images. This chapter proposes deep learning based solutions to automatically quantify knee OA severity from the localised and extracted knee joint images. The objective is to develop and train CNNs to quantify knee OA severity based on the KL grades (0–4). Classification and regression are used to predict the KL grades on an ordinal scale and on a continuous scale. Also, ordinal regression is investigated in an attempt to improve the quantification of knee OA severity.

Previous work on automated assessment of knee OA severity approached it as an image classification problem [2, 15, 29–31]. Previous methods have tested many hand-crafted features based on pixel statistics, textures, edge and object statistics, and transforms [1, 2, 4, 9, 29, 30]. Many classifiers such as the SVM [30], the k-nearest neighbour classifier [29], the weighted neighbour nearest classifier [2, 9], the random forest classifiers [15], and even artificial neural networks (ANN) [31, 32] have been tested for knee image classification. As a baseline (in Section 3.4.2), the state-of-the-art features successful in computer vision tasks such as histogram of oriented gradients [77], local binary patterns [78], and Sobel Gradients [79] are tested. These features are not included in the previous studies to assess knee OA severity. All the previous approaches based on hand-crafted features give low multi-class classification accuracy when classifying knee images and in particular classifying fine-grained successive knee OA grades remains a challenge. As a baseline, the state-of-the-art CNNs features (in Section 3.4.3) are also tested for knee images classification on a small baseline data set from OAI and this approach gave promising results. Motivated by this, the use of CNNs are investigated for quantifying knee OA severity in this chapter.

First, the use of off-the-shelf CNNs are investigated for quantifying knee OA severity through classification and regression. Two approaches are followed for this: 1) using a pre-trained CNN for fixed feature extraction, 2) fine-tuning pre-trained CNN following a transfer learning approach. WNDCHRM, an open source utility for medical image classification [2, 9, 34] is used for benchmarking the classification results obtained from the proposed methods,

Next, three new methods are investigated to automatically quantify knee OA: 1) training a CNN from scratch for multi-class classification of knee OA images; 2) training a CNN to optimise a weighted ratio of two loss functions categorical cross-entropy for multi-class classification and mean-squared error for regression; 3) training a CNN for ordinal regression of knee OA images. The results from these methods are compared to the previous methods. The classification results using both manual and automatic localisation of knee joints are also compared.

The remainder of this chapter is structured as follows: Section 5.2 presents the classification and regression of knee OA images using CNNs that are fine-tuned through transfer learning. Section 5.3 introduces the training of CNNs from scratch for classifying knee OA images and analyses the classification results. Section 5.4 elaborates on the joint training of CNNs for simultaneous classification and regression of knee OA images, and shows the results of joint training. Section 5.5 describes the development and training of a CNN for ordinal regression using a custom loss function. Section 5.6 compares and analyses the results from the four approaches to quantify knee OA severity. Section 5.7 summarises this chapter and presents the conclusions.

5.2 Off-the-shelf CNNs

The use of well-known off-the-shelf CNNs such as the VGG-16 network [63], and comparatively simpler networks like VGG-M-128 network [97], and BVLC reference CaffeNet [98, 99] (which is very similar to the widely-used *AlexNet* model [22]) are investigated to classify knee OA images. These networks are pre-trained for general image classification using a very large dataset: the ImageNet LSVRC dataset [100] which contains more than 1.2 million images in 1000 classes. Initially, features are extracted from the convolutional, pooling, and fully-connected layers of VGG-16, VGG-M-128, and BVLC CaffeNet, and train linear SVMs to classify knee OA images.

The pre-trained networks are fine-tuned for knee OA images classification motivated by the transfer learning approach [76]. Transfer learning is adopted as the OAI dataset is small, containing only a few thousand images. In transfer learning, a base network is first trained on external data, and then the weights of the initial n layers are transferred to a target network [76]. The new layers of the target network are randomly initialised following the Xavier weight initialisation procedure [101]. The random weights initialisations increase the likelihood of the training algorithms during the backpropagation to obtain a global solution through the gradient descent instead of settling to a nearest local solution.

Intuitively, the lower layers of the networks contain more generic features such as edge or texture detectors useful for multiple tasks, whilst the upper layers progressively focus on more task specific cues [76, 99]. This approach is used for both classification and regression, adding new fully-connected layers and backpropagation is used to fine-tune the weights for the complete network on the target loss.

5.2.1 Dataset

The data used for the experiments are knee radiographs taken from the baseline cohort of OAI dataset containing $4,476$ participants. In the entire cohort, Kellgren & Lawrence (KL) grades are available for both knee joints in $4,446$ radiographs and these images are used for this study. The distribution of the knee joint images (in total $8,892$) conditioned on the KL grading scale are: grade 0 - 3433, grade 1 - 1589, grade 2 - 2353, grade 3 - 1222, and grade 4 - 295.

5.2.2 Classification using Pre-trained CNN Features

The VGG-16 network [63] is trained with the OAI dataset. Features are extracted from different layers of the VGG net such as fully-connected (fc7), pooling (pool5), and convolutional (conv5_2) layers to identify the most discriminating set of features. Linear SVMs (LIBLINEAR [102]) are trained with the extracted CNN features for classifying knee OA images, where the ground truth are images labelled with KL grades. Next, the use of simple pre-trained CNNs such as VGG-M-128 [97] and the BVLC CaffeNet model [98] are investigated for classifying the knee OA images. These networks have fewer layers and parameters in comparison to the VGG-16 network. The features are extracted from the fully-connected, pooling, and convolutional layers, using the VGG-M-128 net and the BVLC reference CaffeNet.

Experiments and Results

The knee joint images are split into training ($\sim$70%) and test ($\sim$30%) set based on the distribution of each KL grade. Features are extracted from fully-connected, pooling, and convolution layers of VGG-16, VGG-M-128, and BVLC CaffeNet. Linear SVMs are trained individually for binary and multi-class classifications on the extracted features. WNDCHRM, an open source utility for biological image analysis and medical image classification is used for benchmarking the classification results from the proposed methods in this chapter [2, 9, 34].

WNDCHRM is trained with the same training data so that the classification results from WNDCHRM and CNN features can be compared. The knee OA images are classified in three ways as follows. Classifying healthy knee images (grade 0) with the progressive stages (grade 1, 2, 3, and 4), classifying the images belonging to the successive stages (grade 0 vs 1, grade 1 vs 2, ...) and multi-class classification to classify all the stages of knee OA images.

Table 5.1 shows the test set classification accuracies achieved by WNDCHRM and the CNN features. The CNN features consistently outperform WNDCHRM for classifying healthy knee samples against the progressive stages of knee OA. The features from conv4 layer with dimension 512×13×13 and pool5 layer 256×13×13 of VGG-M-128 net, and conv5 layer with dimension 512×6×6 and pool5 layer with dimension 256×6×6 of BVLC reference CaffeNet give higher classification accuracy in comparison to the fully-connected fc6 and fc7 layers of VGG nets and CaffeNet. Intuitively, the lower layers capture more discriminative low-level features such as edge or shape detectors, and the higher layers tend to contain high-level features specific to object classes as per the training data. Features are also extracted from lower layers such as pool4, conv4_2, pool3, pool2 and train classifiers on top of these features. As the dimension of the bottom layers are high, the training time is increased, however, no improvement in classification accuracy is observed.

In a fine-grained classification task such as knee OA image classification, the accuracy of classifying successive classes tends to be low, as the variations in the progressive stages of the disease are minimal, and only highly discriminant features can capture these variations. From the experimental results, as shown in Table 5.1, the features extracted from CNNs provide significantly higher classification accuracy in comparison to the WNDCHRM, and these features are effective and promising for classifying the consecutive stages of knee OA.

Multi-class classifications are performed using linear SVMs with the CNN features (Table 5.1, multi-class). Again, the CNN features outperform WNDCHRM. The classification accuracies obtained using convolutional (conv4,

Table 5.1: Classification accuracy (%) achieved by the WNDCHRM and pre-trained CNN features.

Category	Classification	WNDCHRM	VGG-16 Net			VGG-M-128 Net			BVLC ref CaffeNet		
			fc7	pool5	conv5_2	fc6	pool5	conv4	fc7	pool5	conv5
Progressive	Grade 0 vs Grade 1	51.5	56.3	61.3	63.5	56.5	63.2	**64.7**	62.0	64.3	63.3
	Grade 0 vs Grade 2	62.6	68.6	74.3	76.7	67.8	75.5	**77.6**	69.6	73.6	73.9
	Grade 0 vs Grade 3	70.6	86.4	91.4	92.4	88.5	90.2	**92.9**	87.9	92.5	91.5
	Grade 0 vs Grade 4	82.8	98.1	98.6	99.3	98.8	99.3	99.2	98.5	**99.4**	99.1
Successive	Grade 1 vs Grade 2	48.8	60.0	64.7	67.3	57.9	63.5	65.3	61.2	**65.8**	62.8
	Grade 2 vs Grade 3	54.5	69.8	76.4	77.0	73.0	77.3	**79.0**	70.3	78.1	77.1
	Grade 3 vs Grade 4	58.6	85.2	88.8	90.0	85.0	90.4	91.2	87.4	**91.6**	91.4
Multi-class	Grade 0 to Grade 2	39.9	51.1	53.4	56.9	51.1	55.0	**57.4**	51.1	54.8	54.4
	Grade 0 to Grade 3	32.0	44.6	48.7	53.9	45.4	50.2	**53.3**	46.9	51.6	50.2
	Grade 0 to Grade 4	28.9	42.6	47.6	53.1	43.8	49.5	**53.4**	44.1	50.8	50.0

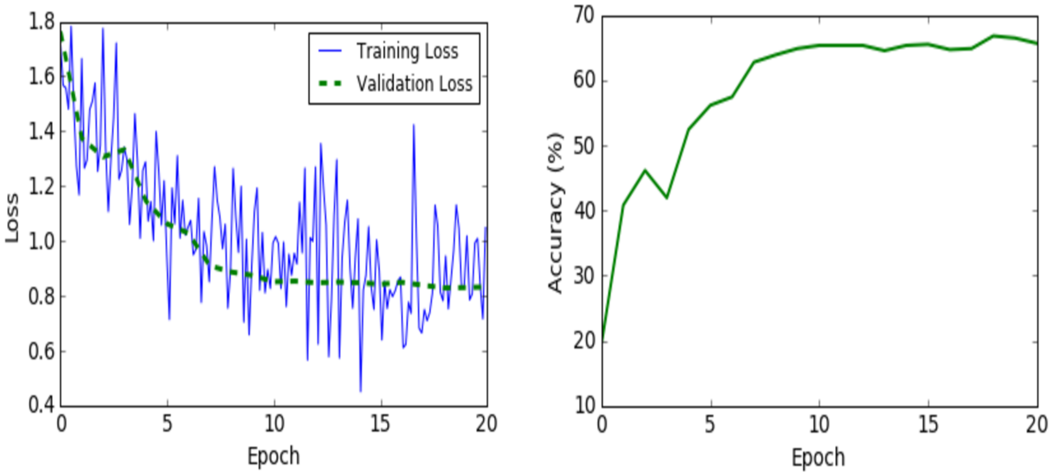


Figure 5.1: Learning curves:training and validation losses (left), and validation accuracy (right) during fine-tuning.

conv5) and pooling (pool5) layers are slightly higher in comparison to fully-connected layer features. There are minimal variations in classification accuracy obtained with the features extracted from VGG-M-128 net and BVLC reference CaffeNet in comparison to VGG-16.

5.2.3 Classification using Fine-tuned CNNs

As a next approach, the BVLC CaffeNet [98] and VGG-M-128 [97] networks are fine-tuned to classify knee images. These two smaller networks are chosen because they contain fewer layers and parameters (∼62M), over the much deeper VGG-16, which has ∼138M parameters. The top fully-connected layer of both networks is replaced and the model is retrained on the OAI dataset using backpropagation. The lower-level features in the bottom layers are also updated during fine-tuning. Standard softmax loss is used as the objective for classification, and accuracy layers are added to monitor the training progress. A Euclidean loss layer (mean squared error) is used for the regression experiments.

Table 5.2: Classification accuracy (%) achieved with the features extracted from fine-tuned BVLC Net.

Classification	Before Fine-Tuning			After Fine-Tuning		
	fc7	pool5	conv5	fc7	pool5	conv5
grade 0 vs grade 1	62.0	64.3	63.3	63.3	**64.3**	61.9
grade 0 vs grade 2	69.6	73.6	73.9	76.3	**77.2**	74.1
grade 0 vs grade 3	87.9	92.5	91.5	**96.7**	96.0	96.3
grade 0 vs grade 4	98.5	99.4	99.1	**99.8**	99.7	99.7
grade 1 vs grade 2	61.2	65.8	62.8	63.3	**66.7**	62.7
grade 2 vs grade 3	70.3	78.1	77.1	**85.8**	83.9	83.3
grade 3 vs grade 4	87.4	91.6	91.4	**94.4**	93.6	92.6
grade 0 to grade 2	51.1	54.8	54.4	**57.4**	57.0	52.0
grade 0 to grade 3	46.9	51.6	50.2	**57.2**	56.5	51.8
grade 0 to grade 4	44.1	50.8	50.0	**57.6**	56.2	51.8

Experiments and Results

Table 5.2 shows the multi-class classification results for the fine-tuned BVLC CaffeNet. The VGG-16 network is omitted in these experiment since the variation in accuracy among the pre-trained CNNs is small, and fine-tuning VGG-16 is more computationally expensive.

The dataset is split into training (60%), validation (10%) and test (30%) sets for fine-tuning. The right-left flipped knee joint images are included in the training set to increase the number of training samples. The networks are fine-tuned for 20 epochs using a learning rate of 0.001 for the transferred layers, and 0.01 for the newly introduced layers. The performance of fine-tuned BVLC CaffeNet is slightly better than VGG-M-128. Hence, the results of fine-tuning BVLC CaffeNet is only shown here. Figure 5.1 shows the learning curves for training and validation loss, and validation accuracy. The decrease in loss and increase in accuracy shows that the fine-tuning is effective and makes the CNN features more discriminative, which improves classification accuracy (Table 5.1). The features extracted from the fully connected (fc7) layer provide slightly better classification in comparison to pooling (pool5) and convolution (conv5) layers.

5.2.4 Regression using Fine-tuned CNNs

Existing work on automatic assessment of knee OA severity treats it as an image classification problem, assigning each KL grade to a distinct category [2,7,14,15]. To date, evaluation of automatic KL grading algorithms has been based on binary and multi-class classification accuracy with respect to these discrete KL grades [1,2,14]. Nevertheless, KL grades are not categorical, but rather represent an ordinal scale of increasing severity. Treating them as categorical during evaluation means that the penalty for incorrectly predicting that a subject with grade 0 OA has grade 4 is the same as the penalty for predicting that the same subject has grade 1 OA. Clearly the former represents a more serious error, yet this is not captured by evaluation measures that treat grades as categorical variables. In this set up, permuting the ordering of the grades has no effect on classification performance. Moreover, the quantisation of the KL grades to discrete integer levels is essentially an artefact of convenience; the true progression of the disease in nature is continuous, not discrete.

The author proposes that it is more appropriate to measure the performance of an automatic knee OA severity assessment system using a continuous evaluation metric like mean squared error. Such a metric appropriately penalises errors in proportion to their distance from the ground truth, rather than treating all errors equally. Directly optimising mean squared error on a training set also naturally leads to the formulation of knee OA assessment as a standard regression problem. Treating it as such provides the model with more information on the structure and relationship between training examples with successive KL grades. It is demonstrated that the use of regression reduces both the mean squared error and improves the multi-class classification accuracy of the model.

The pre-trained BVLC CaffeNet model is fine-tuned using both classification loss (cross entropy on softmax outputs) and regression loss (mean squared error) to compare their performance in assessing knee OA severity. In both cases, the fully connected layer fc7 is replaced with a randomly initialised layer and fine-tuned

Table 5.3: MSE for classification and regression.

Classes	WNDCHRM	CNN-Clsf	CNN-Reg	CNN-Reg*
grade 0 to 4	2.459	0.836	**0.504**	0.576

for 20 epochs, selecting the model with the highest validation performance. The classification network uses a 5D fully connected layer and softmax following the fc7 layer, and the regression network uses a 1D fully connected node with a linear activation.

The models are compared using both mean squared error (MSE) and standard multi-class classification metrics. The mean squared error is calculated using the standard formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2,$$

where n is the number of test samples, y_i is the true (integer) label and $\hat{y_i}$ is the predicted label. For the classification network the predicted labels y_i are integers and for the regression network they are real numbers. A configuration is tested, where the real outputs are rounded from the regression network to produce integer labels. Table 5.3 shows the MSE for classification using the WNDCHRM and the CNN trained with classification loss (CNN-Clsf), regression loss (CNN-Reg), and regression loss with rounding (CNN-Reg*). Regression loss clearly achieves significantly lower mean squared error than both the CNN classification network and the WNDCHRM features.

To demonstrate that the regression loss also produces better classification accuracy, the classification accuracy from the network trained with classification loss and the network trained with regression loss and rounded labels are compared. Rounding, in this case is necessary to allow the use of standard classification metrics. Table 5.4 compares the resulting precision, recall, and F_1 scores. The multi-class (grade 0–4) classification accuracy of the network fine-tuned with

Table 5.4: Comparison of classification performance using classification (left) and regression (right) losses.

Classification	Classification Loss			Regression Loss		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.53	0.64	0.58	0.57	0.92	0.71
1	0.25	0.19	0.22	0.32	0.14	0.20
2	0.44	0.32	0.37	0.71	0.46	0.56
3	0.37	0.47	0.41	0.78	0.73	0.76
4	0.56	0.54	0.55	0.89	0.73	0.80
Mean	0.43	0.44	0.43	0.61	0.62	0.59

regression loss is 59.6%. The network trained using regression loss clearly gives superior classification performance. The author suspects this is due to the fact that using regression loss gives the network more information about the ordinal relationship between the KL grades, allowing it to converge on parameters that better generalise to unseen data.

5.2.5 Discussion

The initial approach to quantify knee OA severity used features extracted from pre-trained CNNs. Three pre-trained networks are investigated and it is found that the BVLC reference CaffeNet and VGG-M-128 networks perform best. A linear SVM trained on features from these networks achieved significantly higher classification accuracy (53.4%) in comparison to the previous state-of-the-art (28.9%). The features from pooling and convolutional layers were found to be more accurate than the fully connected layers. Fine-tuning the networks by replacing the top fully connected layer gave further improvements in multi-class classification accuracy.

Previous studies have assessed their algorithms using binary and multi-class classification metrics. The author proposes that it is more suitable to treat KL grades as a continuous variable and assess accuracy using mean squared error. This approach allows the model to be trained using regression loss so that errors are penalised in proportion to their severity, producing more accurate predictions. This approach also has the nice property that the predictions can fall between grades, which aligns with continuous disease progression.

In summary, this section presented two approaches based on the existing pre-trained CNNs for quantifying knee OA severity: first, the CNNs are used for fixed feature extraction and next, the CNNs are fine-tuned using transfer learning. Both the approaches outperformed the previous state-of-the-art, the WNDCHRM classifier giving promising results. As a next logical step, CNNs are trained from scratch to investigate if this leads to further improvement in quantifying knee OA severity.

5.3 Training CNNs from scratch

Training a CNN from scratch (or full training) is challenging and complicated, because it requires a large amount of annotated training data. The learning curves during training should ensure proper convergence to generalise well avoiding overfitting [23]. An alternative to full training is transfer learning, fine-tuning CNNs pre-trained in other domain (for instance ImageNet dataset with natural images) to a target domain for instance medical domain. However, the knowledge transfer may be limited by the substantial differences between the source and the target domains, which may mitigate the performance of the fine tuned CNNs. Nevertheless, with sufficient labelled training data and carefully selected hyper-parameters, fully trained CNNs can outperform fine-tuned CNNs and hand-crafted alternatives [20, 23].

Fully trained CNNs have been found to be highly successful in many medical applications [20, 23]. Some of the applications that use fully trained CNNs for musculo-skeletal (including knee) image analysis are knee cartilage segmentation using multi-stream CNNs [24], total knee arthoplasty kinematics by real-time 2D/3D registration using CNN regressors [103], automated skeletal bone age assessment in X-ray images using deep learning [104], posterior-element fractures detection on spine CT using deep convolutional networks [105], and automated anatomical landmarks detection on distal femur bone on 3D image analysis using CNNs. Motivated by these approaches, CNNs are trained from scratch to quantify knee OA severity using both classification and regression.

5.3.1 Dataset and Preprocessing

The data used for the initial experiments are taken from the baseline OAI dataset. There are 4,446 X-ray images with the KL grade annotations in this dataset. The MOST dataset is included for later experiments and this dataset consists of 2,920 X-ray images with KL grade annotations. Two set of knee joint images are used separately for the experiments: 1) extracted after automatic localisation and 2) extracted after manual annotation of the ROI. This is to compare the quantification performance of the CNNs trained with knee joints from automatic localisation and manual annotation. As a preprocessing step, all the knee joint images are subjected to histogram equalisation for intensity level normalisation. The images were resized to 256×256 pixels for the initial experiments. Later, the input image size is changed to 200×300. This size is chosen to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints. Right-left flip of the knee joint images are used to generate more training data.

5.3.2 Initial Configuration

A CNN is configured with a lightweight architecture with 4 layers of learned weights: 3 convolutional layers and 1 fully connected layer. As the training data set is relatively small, a lightweight architecture is considered with minimal (4.5 million) parameters in comparison to the existing CNNs. Table 5.5 shows the CNN configuration in detail. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. A max pooling layer is included after each convolution stage. The final pooling layer is followed by a fully connected layer

Table 5.5: Initial CNN configuration.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11 ×11	2	32×128×128
maxPool1	–	3×3	3	32×42×42
conv2	96	7×7	1	96×42×42
maxPool2	–	3×3	3	96×14×14
conv3	128	3×3	1	128×14×14
maxPool3	–	3×3	2	128×4×4
fc4	–	–	–	2048
fc5	–	–	–	5

(fc4), and a softmax dense layer (fc5) with an output shape 5 for the multi-class classification of (0–4) ordinal KL grades. A drop out layer with a drop out ratio of 0.5 is included after the fully connected layer (fc5) to avoid overfitting. The input images are of size 256×256 pixels and fed to the network after sub-sampling by a factor of 2. So, the input size is 128×128 pixels.

5.3.3 Training Process and Initial Results

The network parameters are trained from scratch with the knee joint images as training samples and the KL grades (0, 1, 2, 3 or 4) as labels. To start, the knee joint images extracted manually from the radiographs of the OAI dataset are used. The dataset is split into training (70%) and test (30%) sets. The validation (10%) data is taken from the training set. The network is trained to minimise categorical cross entropy for multi-class classification. *Stochastic gradient descent* (SGD) is used with default parameters: decay = $1e^{-6}$, momentum = 0.9, and nesterov = True and the initial learning rate is set to 0.0001. The networks are trained with fixed learning rate in the initial experiments. The Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$ is tested, instead of SGD for the later experiments. The benefits of the Adam optimiser are that it uses adaptive learning rates and provides faster convergence.

This network achieves a multi-class classification accuracy of 44.7% on the test data. The mean-squared error is 1.75. Table 5.6 shows the classification results:

Table 5.6: Classification results of the initial CNN configuration.

grade	Precision	Recall	F_1 Score
0	0.45	0.92	0.60
1	0.24	0.07	0.11
2	0.49	0.18	0.26
3	0.50	0.39	0.44
4	1.00	0.01	0.02
Mean	0.45	0.45	0.37

precision, recall, and F_1 score of the initial configuration. The results show that the classification performance is low and the mean-squared error is high. These are initial results and the hyper-parameters of this network are tuned to improve the classification performance. Further, the number of convolutional layers, convolutional-pooling stages, the number of convolutional kernels, kernel sizes and other parameters are experimented.

The terms 'parameters' and 'hyper-parameters' in machine learning are often used interchangeably, but there is a difference between them. Parameters are learned by a classifier or a machine learning model from the training data, for instance weights or coefficients of the independent variables. Hyper-parameters are the settings used to optimise the performance of a classifier or a model and they are not fit based on the training data. The hyper-parameters for a CNN include the number and size of the hidden layers, learning rate and its decay, drop out regularisation, gradient clipping threshold and other settings.

5.3.4 Tuning Hyper-parameters

After the initial CNN configuration giving low classification accuracy (44.7%), as a first step the depth of the network is increased. A convolutional layer and a pooling layer are included. This increases the number of layers with learned weights to 5 layers: 4 convolutional layers and 1 fully connected layer. SGD with default parameters: decay = $1e^{-6}$, momentum = 0.9, and nesterov = True, is used for training this network. Learning rates from 0.0001 to 0.01 with an incremental

Table 5.7: CNN architecture (CNN-1) after tuning hyper-parameters.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×128×128
maxPool1	–	3×3	2	32×63×63
conv2	96	5×5	1	64×63×63
maxPool2	–	3×3	2	64×31×31
conv3	128	3×3	1	128×31×31
maxPool3	–	3×3	2	128×15×15
conv4	256	3×3	1	256×15×15
maxPool4	–	3×3	2	256×7×7
fc5	–	–	–	1024
fc6	–	–	–	5

increase by a factor 10 are tested, and the learning rate 0.001 is found to be the best. After experimenting with the convolutional kernel size, the number of kernels in the convolutional layer, the number of outputs of the fully connected layer and other parameters, the final architecture in this configuration is obtained. Table 5.7 shows the CNN architecture in detail.

After 20 epochs of training, this network gave a multi-class classification accuracy of 55.2% with a mean-squared error 0.803 on the validation data. After 35 epochs the network achieves the best results for this configuration with a classification accuracy of 60.4% and mean-squared error 0.838. Table 5.8 shows the classification results: precision, recall, and F_1 score of this network. There is an improvement in the overall classification results in comparison to the previous results (Table 5.6). Figure 5.2 shows the learning curves with increase in the training and validation accuracies, and decrease in the training and validation losses whilst training this network. It can be observed from the learning curves (Figure 5.2), after 32 epochs there is an increase in validation loss with decrease in training loss and also there is no further increase in validation accuracy whilst training accuracy increases: the network is starting to overfit. A drop out regularisation by a ratio of 0.5 is included after the fully connected layer (fc5) to mitigate overfitting. Also, data augmentation is used to increase the training samples by including the right-left flip of the knee

Table 5.8: Classification results after tuning hyper-parameters.

grade	Precision	Recall	F_1 Score
0	0.57	0.90	0.70
1	0.31	0.11	0.16
2	0.64	0.45	0.53
3	0.74	0.77	0.76
4	0.86	0.72	0.78
Mean	0.58	0.60	0.57

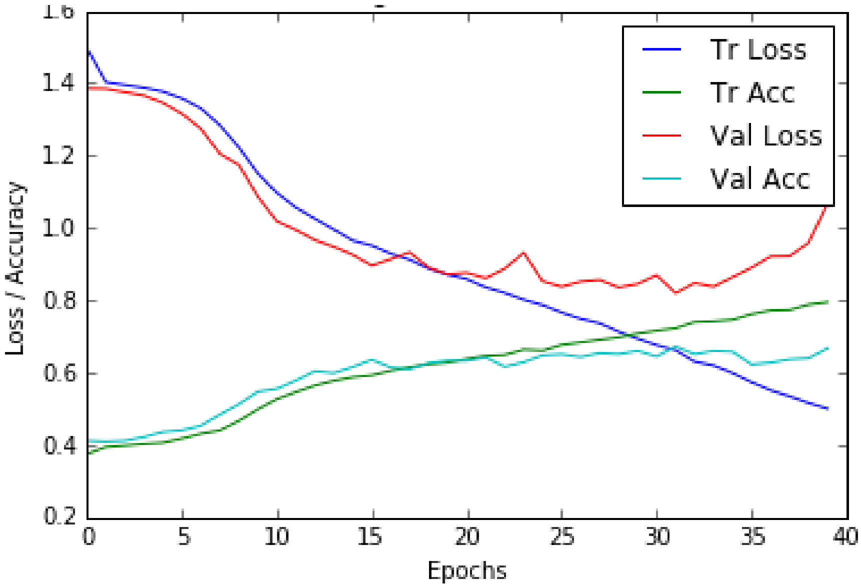


Figure 5.2: Learning curves: training and validation losses, and accuracies of the fully trained CNN

joints and this doubles the number of training samples. Drop out regularisation after convolutional layers and fully connected layers, and l2-norm weight regularisations are used to further mitigate overfitting in the next set of experiments.

Next, the depth of the network is further increased by increasing the number of layers with learned weights, continuing the experimentation with the other associated hyper-parameters. Up to 5 convolutional-pooling stages followed by two fully connected layers are tested. The classification accuracy with 4 convolutional-pooling stages is 60.8% and with 5 convolutional-pooling stages is 61%.

Table 5.9: CNN architecture (CNN-2) after tuning hyper-parameters.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×128×128
maxPool1	–	3×3	2	32×63×63
conv2	64	5×5	1	64×63×63
maxPool2	–	3×3	2	64×31×31
conv3-1	64	3×3	1	64×31×31
conv3-2	64	3×3	1	64×31×31
maxPool3	–	3×3	2	64×15×15
conv4-1	96	3×3	1	96×15×15
conv4-2	96	3×3	1	96×15×15
maxPool4	–	3×3	2	96×7×7
fc5	–	–	–	1024
fc6	–	–	–	5

Previous networks use a single convolutional layer followed by a pooling layer. Next, cascaded convolutional layers are used in a convolution-pooling stage like VGG-16 model. Each convolutional layer is followed by a ReLU activation. Figure 5.9 shows the CNN architecture that gives the best results in this approach. This network gives a classification accuracy of 60.1% with a mean-squared error 0.838.

Inspired by the success of VGG networks [63], a network with cascaded convolutional layers of uniform (3×3) kernel size and (2×2) max pooling with stride 2 is trained, and the hyper-parameters are tuned. This network gives a classification accuracy of 57.5% with a mean-squared error 0.961. There is no further improvement in the classification results in comparison to the previous results.

5.3.5 Training Off-the-shelf CNNs from Scratch

Earlier, the widely used off-the-shelf CNNs such as BVLC reference CaffeNet [98, 99] (which is very similar to the AlexNet model [22]), VGG-M-128 network [97], and VGG-16 network [63] were fine-tuned for knee images classification. The pre-trained VGG-16 network has ∼138 million free parameters, and the other networks, Alexnet with ∼62 million and the VGG-M-128 with ∼26 million parameters, are

Table 5.10: AlexNet architecture.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	96	11×11	4	96×64×64
maxPool1	–	3×3	2	96×31×31
conv2	256	5×5	1	256×31×31
maxPool2	–	3×3	2	256×15×15
conv3	384	3×3	1	384×15×15
conv4	384	3×3	1	384×15×15
conv5	256	3×3	1	256×15×15
maxPool5	–	3×3	2	256×7×7
fc6	–	–	–	4096
fc7	–	–	–	4096
fc8	–	–	–	5

relatively simple. Training these networks, in particular VGG-16, from scratch is computationally very expensive due to the depth and the number of free parameters. Previously trained CNNS have relatively fewer parameters (∼4 to 6 million) to suit the relatively small dataset with a few thousand of training examples.

Next, CNNs are fully trained using the AlexNet and the VGG-M-128 architectures. This is to compare the classification performance of these networks to the previously trained networks from scratch. Table 5.10 shows the AlexNet architecture in detail. The convolutional layers conv1 and conv2 in this network are followed by Relu and batch normalisation layers. The two fully connected layers (fc6) and (fc7) are followed by a drop out regularisation by a ratio 0.5. This network was pre-trained for 1,000 classes in the ImageNet [33] dataset. The output of the last fully connected layer (fc8) is replaced with a 5 output dense layer for multi-class knee OA image classification. This network is trained using SGD with default parameters. Learning rates from 0.00001 to 0.01 with an incremental increase by a factor 10 are tested. The learning rate set at 0.001 gives the best results.

The fully trained AlexNet gives a classification accuracy of 57.2% with a mean-squared error 0.741. Table 5.11 shows the classification results; precision, recall, and F_1 score of the fully trained AlexNet model. These results show that

Table 5.11: Classification results of the fully trained AlexNet.

grade	Precision	Recall	F_1 Score
0	0.65	0.61	0.63
1	0.29	0.36	0.32
2	0.59	0.55	0.57
3	0.75	0.73	0.74
4	0.77	0.79	0.78
Mean	0.59	0.57	0.58

the classification accuracy achieved by the fully trained AlexNet is low (57.2%) in comparison to the accuracy (60.8%) achieved by previous networks. Moreover, this network is overfitting. This is evident from the learning curves (Figure 5.3) obtained whilst training this network. After 30 epochs, the learning curves show an increase in validation loss whilst the training loss is decreasing and there is no improvement in the validation accuracy whilst the training accuracy keeps increasing. The reason for overfitting is the number of training samples in the dataset (∼10,000) is very low in comparison to the number of free parameters (∼62 million) in AlexNet. This model was originally developed and trained on datasets like ImageNet [33] that consists of more than ∼1.2 million images. There are two fully connected layers with 4,096 outputs in the AlexNet and these layers contribute to more than 95% of the total free parameters in this network. Next, a relatively simple architecture (VGG-M-128) is investigated for the knee OA images classification.

The VGG-M-128 network is a simplified model of the AlexNet [63]. The last fully connected layer (fc7) of AlexNet has 4,096 outputs. The number of fc7 outputs is reduced to 128 in VGG-M-128. This reduces the number of free parameters and this network contains (∼26 million) parameters in total. The AlexNet configuration is retained in the VGG-M-128 network with a few changes in the architecture. The kernel size of the first convolutional layer is reduced to (7×7) and the stride is reduced to 2. The number of filters is fixed to 512 in the conv3, conv4, and conv5 layers. Table 5.12 shows the architecture details. This network parameters are

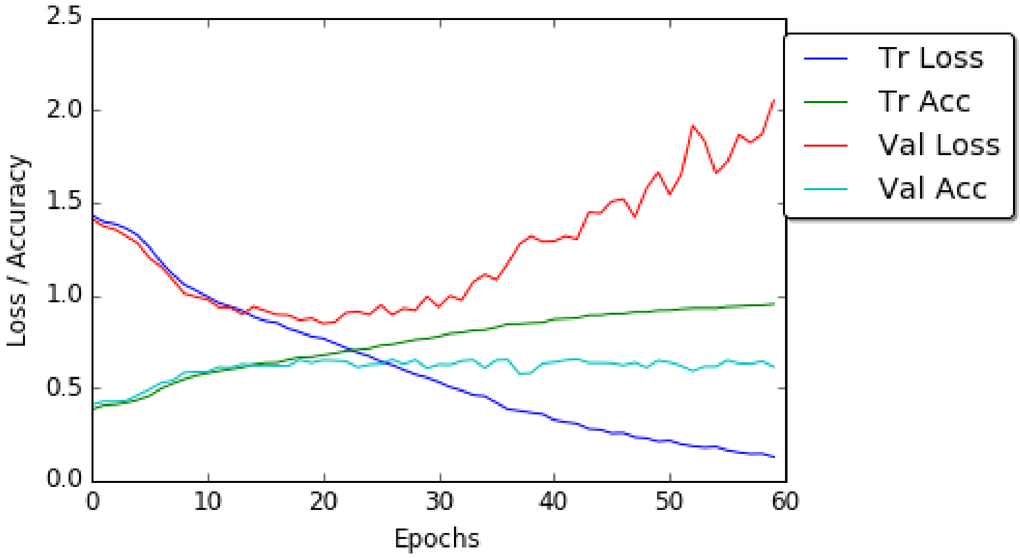


Figure 5.3: Learning curves: training and validation losses, and accuracies of the fully trained AlexNet.

trained from scratch using SGD with default parameters: decay = 1e^{-6}, momentum = 0.9, and nesterov = True. The learning rate is fixed to 0.001 after testing different rates like before.

This network gives a classification accuracy of 56.3% and the mean-squared error is 0.685. Table 5.13 shows the classification results of this network. The results show a slightly lower classification accuracy (56.3%) in comparison to the previous results. There is no significant difference in the precision, recall, and F_1 score of this network in comparison to the AlexNet classification results (Table 5.11). This network is also overfitting like the AlexNet. This is evident from the learning curves (Figure 5.4) of this network. The learning curves show increase in the validation loss after 30 epochs and the validation accuracy remains almost the same. The drop out regularisations after the fully connected layers fc6 and fc7 are not able to fully mitigate overfitting. The reason for overfitting remains the same as for AlexNet. The number of training samples is very low even for the number of free parameters in this network (∼26 million).

Table 5.12: VGG-M-128 architecture.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	96	7×7	2	96×128×128
maxPool1	–	3×3	2	96×63×63
conv2	256	5×5	1	256×32×32
maxPool2	–	3×3	2	256×15×15
conv3	512	3×3	1	512×15×15
conv4	512	3×3	1	512×15×15
conv5	512	3×3	2	512×8×8
maxPool5	–	3×3	2	512×3×3
fc6	–	–	–	4096
fc7	–	–	–	128
fc8	–	–	–	5

Table 5.13: Classification results of the fully trained VGG-M-128.

grade	Precision	Recall	F_1 Score
0	0.66	0.65	0.66
1	0.27	0.42	0.33
2	0.62	0.46	0.53
3	0.77	0.69	0.72
4	0.87	0.73	0.79
Mean	0.60	0.56	0.58

5.3.6 Best Performing CNN for Classification

After experimenting with different configurations, the network in Table 5.14 is found to be the best for classifying knee images. This network is similar to the previous configuration (Table 5.9), but with slight variations. The network contains five layers of learned weights: four convolutional layers and a fully connected layer. The total number of free parameters in the network is ∼5.4 million. Each convolutional layer in the network is followed by batch normalisation and a ReLU activation layer. After each convolutional stage there is a max pooling layer. The final pooling layer (maxPool4) is followed by a fully connected layer (fc5) and a softmax dense (fc6) layer. To avoid overfitting, a drop out layer with a drop out ratio of 0.25 is included after the last convolutional (conv4) layer and a drop out layer with a drop out ratio

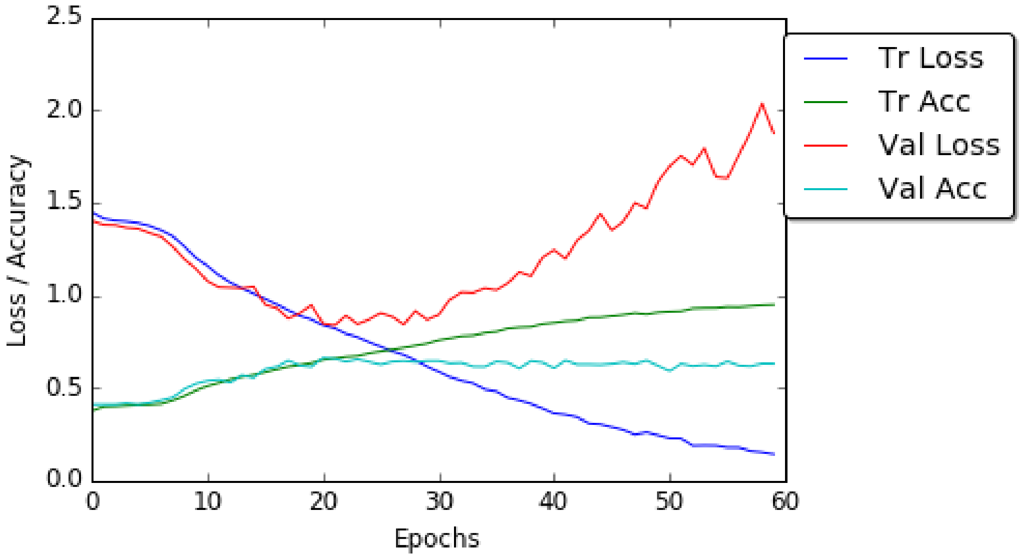


Figure 5.4: Learning curves: training and validation losses, and accuracies of the fully trained VGG-M-128 network.

of 0.5 after the fully connected layer (fc5). Also, a L2-norm weight regularisation penalty of 0.01 is applied in the last two convolutional layers (conv3 and conv4) and the fully connected layer (fc5). Applying a regularisation penalty to other layers increases the training time whilst not introducing significant variation in the learning curves. The network is trained to minimise categorical cross-entropy loss using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\mathrm{e}^{-8}$. The inputs to the network are knee images of size 200×300. This size is selected to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints.

First, this network is trained using the OAI dataset like the previous network trainings. This network achieves a classification accuracy of 61% with a mean-squared error 0.861. Next, training samples are included from the MOST dataset. This network achieves a classification accuracy of 61.8% with a mean-squared error 0.735 for the combined OAI-MOST dataset. There is a slight increase in the classification accuracy (0.8%) and decrease in the mean-squared error (0.126). Table 5.15 shows the classification results: precision, recall, and F_1

Table 5.14: Best performing CNN for classifying the knee images.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2	64	5×5	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8
fc5	–	–	–	1024
fc6	–	–	–	5

Table 5.15: Classification results of the best performing fully trained CNN.

grade	Precision	Recall	F_1 Score
0	0.65	0.83	0.73
1	0.30	0.10	0.14
2	0.51	0.60	0.55
3	0.77	0.69	0.73
4	0.87	0.70	0.78
Mean	0.59	0.62	0.59

score of this network for the combined OAI-MOST dataset. Figure 5.5 shows the learning curves whilst training this network. The learning curves show proper convergence of the training and validation losses with consistent increase in the training and validation accuracies till they reach constant values.

To sum up, a high classification accuracy (61%) is achieved with the CNN (Table 5.14) trained from scratch and outperform the VGG-M-128 and the AlexNet trained from scratch. The fully trained AlexNet gives a classification accuracy of 57.2% and VGG-M-128 gives an accuracy of 56.3%. The classification results of the methods proposed in this section and the previous state-of-the-art are compared in the next section.

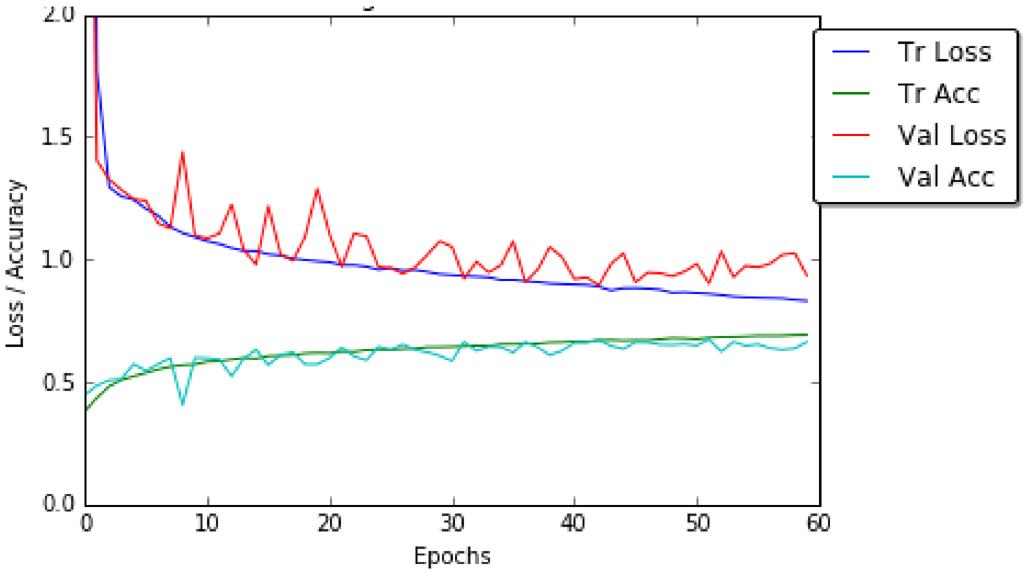


Figure 5.5: Learning curves: training and validation losses, and accuracies of the best performing fully trained CNN.

Table 5.16: Classification results of the proposed methods and the existing methods.

Method	Test Data	Accuracy	Mean-Squared Error
Wndchrm	OAI & MOST	34.8%	2.112
Fine-Tuned BVLC CaffeNet	OAI	57.6%	0.836
Fully trained CNN	OAI	61%	0.861
Fully trained CNN	OAI & MOST	**61.8%**	**0.735**

5.3.7 Classification Results

The classification results of the fully trained network is compared to WNDCHARM, the multi purpose medical image classifier [9, 14, 34] that gave the previous best results for automatically classifying knee OA X-ray images, and to previous results (Table 5.2) on fine-tuning BVLC reference caffenet for this task (Section 5.2). WNDCHARM is trained with the data taken from the OAI and MOST datasets.

Table 5.16 shows the multi-class classification accuracy and mean-squared error of the fine-tuned BVLC CaffeNet, the network trained from scratch and WND-CHARM for the OAI and MOST datasets. The results show that the network trained from scratch for classifying knee OA images clearly outperforms

WNDCHARM. This shows learning feature representations using CNNs for fine-grained knee OA images classification is highly effective and a better approach in comparison to using a combination of hand-crafted features in WNDCHARM. The other reason for low classification accuracy of WNDCHARM is that it uses only a balanced dataset for training. Both the OAI and MOST datasets are very unbalanced and in particular the number of knee images available in KL grade 4 is very small, ~5% in total.

Moreover, these results show an improvement over previous methods that used fine-tuned off-the-shelf networks such as VGG-M-128 and the BVLC Reference CaffeNet for classifying knee OA X-ray images through transfer learning. These improvements are due to the lightweight architecture of the network trained from scratch with less (~5.4 million) free parameters in comparison to 62 million free parameters of BVLC CaffeNet for the small amount of training data available. The off-the-shelf networks are trained using a large dataset like ImageNet containing millions of images, whereas the dataset used in this experiment contains much fewer (~10,000) training samples. Furthermore, the results show an increase in classification accuracy from 61% to 61.8% when the MOST dataset is included in the training set. This result is promising and it shows that with more training data the CNN performance can further improve. Next, the use of regression by fully trained CNNs is investigated to improve the quantification performance.

5.3.8 Comparison of Manual and Automatic Localisations

The classification results obtained with manually extracted and automatically localised knee joints are compared. The same configuration 5.14 is used to train a CNN with automatically localised knee joints. The previous training-test data split is retained to make valid comparisons. Table 5.17 shows the multi-class classification accuracy and the mean-squared error for the classification using manually extracted and automatically localised knee joints. Table 5.18 shows the classification results: precision, recall, and F_1 score for both the classifications.

Table 5.17: Comparison of classification results obtained using manually annotated knee joints to automatically localised knee joints.

Method	Classification Accuracy	Mean-Squared Error
Manual localisation	**61.8%**	**0.735**
Automatic localisation	61.2%	0.741

Table 5.18: Comparison of manual and automatic localisation performance.

Grade	Manual Localisation			Automatic Localisation		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.65	0.83	0.73	0.66	0.78	0.71
1	0.30	0.10	0.14	0.30	0.21	0.25
2	0.51	0.60	0.55	0.52	0.44	0.48
3	0.77	0.69	0.73	0.71	0.77	0.74
4	0.87	0.70	0.78	0.76	0.86	0.81
Mean	0.59	0.62	0.59	0.58	0.60	0.59

From the results (Table 5.17 and Table 5.18), it is evident that classification of automatically localised knee joint images is on par with the classification of manually extracted knee joints.

5.3.9 Training CNNs for Regression

CNNs are trained from scratch to classify knee images in the previous approach. The outcomes are ordinal KL grades (0, 1, 2, 3 or 4) that quantify knee OA severity. CNNs are trained for regression in the next approach. This is to assess knee OA severity in a continuous scale (0–4). The author has argued earlier (Section 5.2.4) that it is more appropriate to assess knee OA in a continuous scale as knee OA is progressive in nature, not discrete. The existing CNNs are fine-tuned to quantify knee OA severity using regression.

Initial Configuration

A CNN is trained for regression using almost the same architecture (Table 5.14) that gave the highest multi-class classification accuracy previously. The last fully connected layer (fc6) with softmax activation and an output shape 5 for multi-class classification is replaced with a linear activation with an output shape of 1 for regression. The CNN is trained to minimise mean-squared error using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}^{-8}$. Like before, the inputs to the network are images of size 200×300. The data for training is taken from both the OAI and the MOST datasets. Both these datasets contain discrete KL grade (0, 1, 2, 3 or 4) annotations for the knee joints. These labels are used in the previous approach to train classifiers. However, there is no ground truth of KL grades on a continuous scale available for either of these datasets to train a network directly for regression output. Hence, the discrete KL grades are used as labels to train CNNs for regression.

Initial Results

This CNN gives a mean-squared error of 0.654 on the test data after training. In comparison to the mean-squared error achieved by the classifier (0.898) with almost the same architecture, there is definitely an improvement in the quantification using regression. The performance metrics; accuracy, precision, recall, and F_1 score are computed for the regression results by rounding the predicted continuous grade to the next integer value. Rounding, in this case, is necessary to allow the use of standard classification metrics and compare the performances of classification and regression. Table 5.19 shows the precision, recall, and F_1 score for regression after rounding the continuous grades. In comparing these results to the previous classification results (Table 5.15), there is a decrease in precision, recall, and F_1 score. The classification accuracy achieved by regression is 36.9% with a mean-squared error 0.75. From these results it is evident that the regression performance is low in this initial configuration. Next, the hyper parameters of this network are tuned to improve the regression performance.

Table 5.19: Results of the initial network trained for regression after rounding the predicted continuous grades.

Grade	Precision	Recall	F_1 Score
0	0.78	0.18	0.29
1	0.24	0.83	0.37
2	0.49	0.32	0.39
3	0.63	0.42	0.50
4	0.57	0.20	0.30
Mean	0.57	0.37	0.36

Tuning the Hyper-parameters

The experiment is continued by varying the number of layers with learned weights in the architecture, number of convolutional-pooling stages, number of kernels and kernel sizes in the convolutional layers and regularisations to avoid overfitting. The architecture in Table 5.20 is found to be the best for quantifying knee OA severity using regression. This network contains seven layers of learned weights: six convolutional layers and a fully connected layer. This network has ~5.6 million free parameters in total. Each convolutional layer is followed by batch normalisation and a ReLU layer. The last pooling layer (maxPool4) is followed by two dense layers: fc5 with ReLU and fc6 with linear activations. A drop out layer with a drop out ratio 0.5 is added after fc6. The network is trained to minimise the mean-squared error using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}^{-8}$. The network is trained with knee images taken from the OAI and the MOST datasets. Figure 5.6 shows the learning curves: training and validation losses whilst training this network. The learning curves show convergence in the losses.

5.3.10 Comparison of Classification and Regression Results

The best performing CNN for regression gives a mean-squared error of 0.574. After rounding the continuous grade predictions, this network achieves a multi-class classification accuracy of 54.7% and the mean-squared error is 0.661.

Table 5.20: Best performing CNN for regression of the knee images.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×158
maxPool1	–	3×3	2	32×49×74
conv2	64	5×5	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3-1	64	3×3	1	64×24×36
conv3-2	64	3×3	1	64×24×36
maxPool3	–	3×3	2	64×11×17
conv4-1	128	3×3	1	96×11×17
conv4-2	128	3×3	1	96×11×17
maxPool4	–	3×3	2	96×5×8
fc5	–	–	–	1024
fc6	–	–	–	1

Table 5.21: Comparison of classification and regression results.

Method	Accuracy	MSE (before rounding)	MSE (after rounding)
CNN-Classification	**61.8%**	0.735	–
CNN-Regression	54.7%	**0.574**	0.661

Table 5.21 shows the accuracy and mean-squared error for the fully trained CNN for classification and regression. The results show that the multi-class classification accuracy calculated after rounding the output is low for CNN-regression. The main reason for this likely is training the regression network with ordinal labels instead of continuous labels. There is also a decrease in accuracy due to the rounding of regression output and the rounding is necessary to compute standard classification metrics. On the other hand, the mean-squared error of the fully trained CNN for regression is low in both the cases before rounding (0.574) and after rounding (0.661) in comparison to the fully trained CNN for regression. Table 5.22 shows the precision, recall, and F_1 score of the rounded regression output and the classification output. These results show that the network trained with classification loss outperforms the regression loss. The reason for this is again likely the lack of continuous KL grade ground truth to train a CNN directly for regression output.

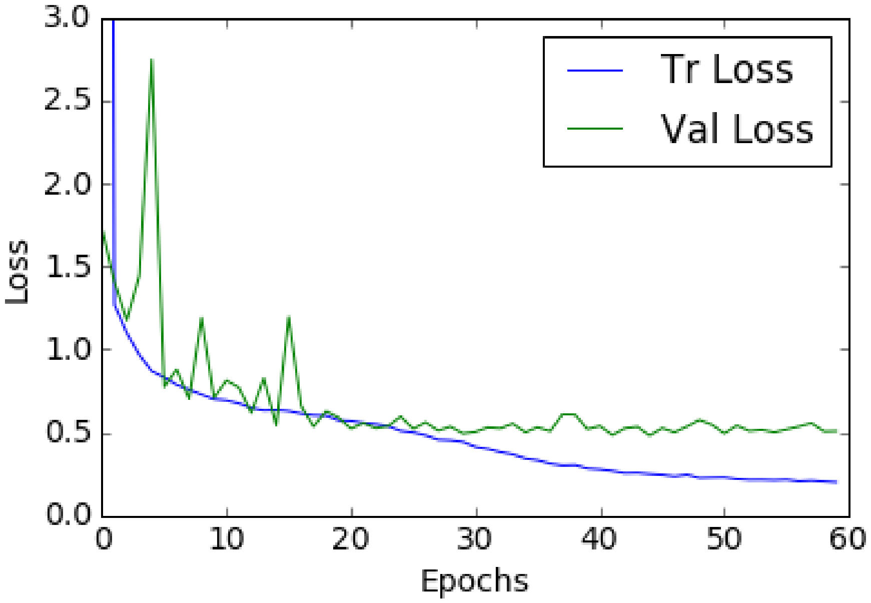


Figure 5.6: Learning curves: training and validation losses for the best performing CNN for regression.

To sum up, training a CNN from scratch for regression output gives low mean-squared error. The lack of ground truth affects the performance of the regression. To overcome this drawback, in the next approach multi-objective convolutional learning is investigated to quantify knee OA severity.

5.4 Multi-objective Convolutional Learning

In general, assessing knee OA severity is based on the multi-class classification of knee images and assigning KL grade to each distinct category [1, 2, 9, 14]. The author argued previously that assigning a continuous grade (0–4) to knee images through regression is a better approach for quantifying knee OA severity as the disease is progressive in nature. However, there is no ground truth i.e. KL grades on a continuous scale to train a network directly for regression output. Therefore, the networks are trained using multi-objective convolutional learning [106–110] to optimise a weighted-ratio of two loss functions: categorical cross-entropy and mean-squared error. Mean squared error gives the network information about

Table 5.22: Comparison of the regression and classification performances.

Grade	Regression			Classification		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.70	0.71	0.70	0.65	0.83	0.73
1	0.29	0.42	0.34	0.30	0.10	0.14
2	0.52	0.39	0.45	0.51	0.60	0.55
3	0.67	0.51	0.58	0.77	0.69	0.73
4	0.58	0.55	0.57	0.87	0.70	0.78
Mean	0.57	0.55	0.55	0.59	0.62	0.59

ordering of grades, and cross entropy gives information about the quantisation of grades. Intuitively, optimising a network with two loss functions provides a stronger error signal and it is a step to improve the overall quantification, considering both classification and regression results.

5.4.1 Initial Configuration

The same architecture of the best performing CNN is used for classification (Table 5.14) as an initial configuration to jointly train a CNN for classification and regression outputs. Table 5.23 and Figure 5.7 shows the configuration details of the initial configuration. The network has five layers with learned weights: four convolutional layers and a fully connected layer. The total free parameters in the network are ~5.4 million. The last fully connected layer (fc5) is followed by two dense layers with softmax and linear activations for simultaneous multi-class classification and regression outputs. Drop out layers with a drop out ratio 0.25 are included after the conv4 layer and a drop out ratio 0.5 after the fc6 layer to avoid overfitting. In addition to this, a L2-norm weight regularisation penalty of 0.01 is applied in conv3, conv4 and fc6 layers to avoid overfitting. Applying a regularisation penalty to other layers did not introduce significant variations in the learning curves. Unlike the previous approaches, this network is trained to minimise a weighted ratio of two loss functions: categorical cross-entropy and mean-squared error. After testing different values from 0.2 to 0.6 for the weight of

Table 5.23: Initial configuration to jointly train a CNN for classification and regression outputs.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2	64	5×5	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3	96	3×3	1	128×24×36
maxPool3	–	3×3	2	128×11×17
conv4	128	3×3	1	256×11×17
maxPool4	–	3×3	2	256×5×8
fc5	–	–	–	1024
fc6-Clsf	–	–	–	5
fc6-Reg	–	–	–	1

regression loss, a ratio of 0.5 is fixed, as this ratio gives optimal results.

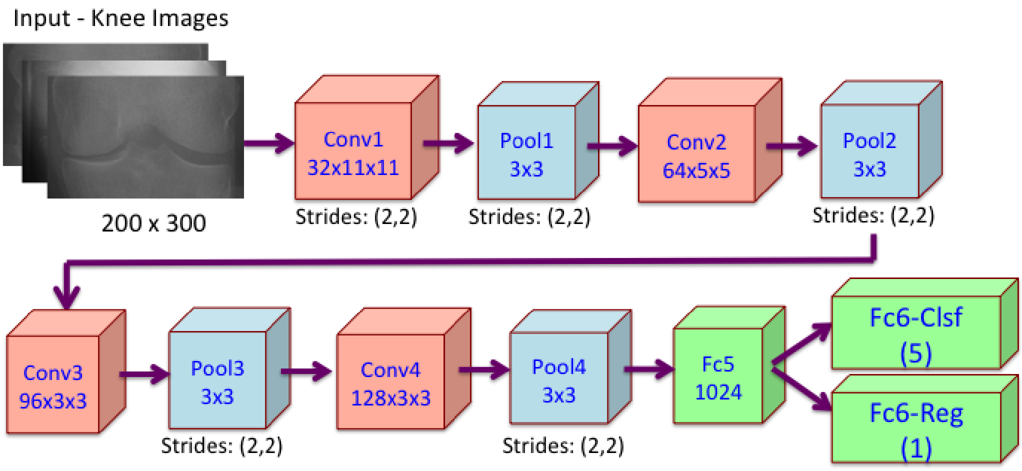


Figure 5.7: Initial configuration to jointly train a CNN for classification and regression outputs.

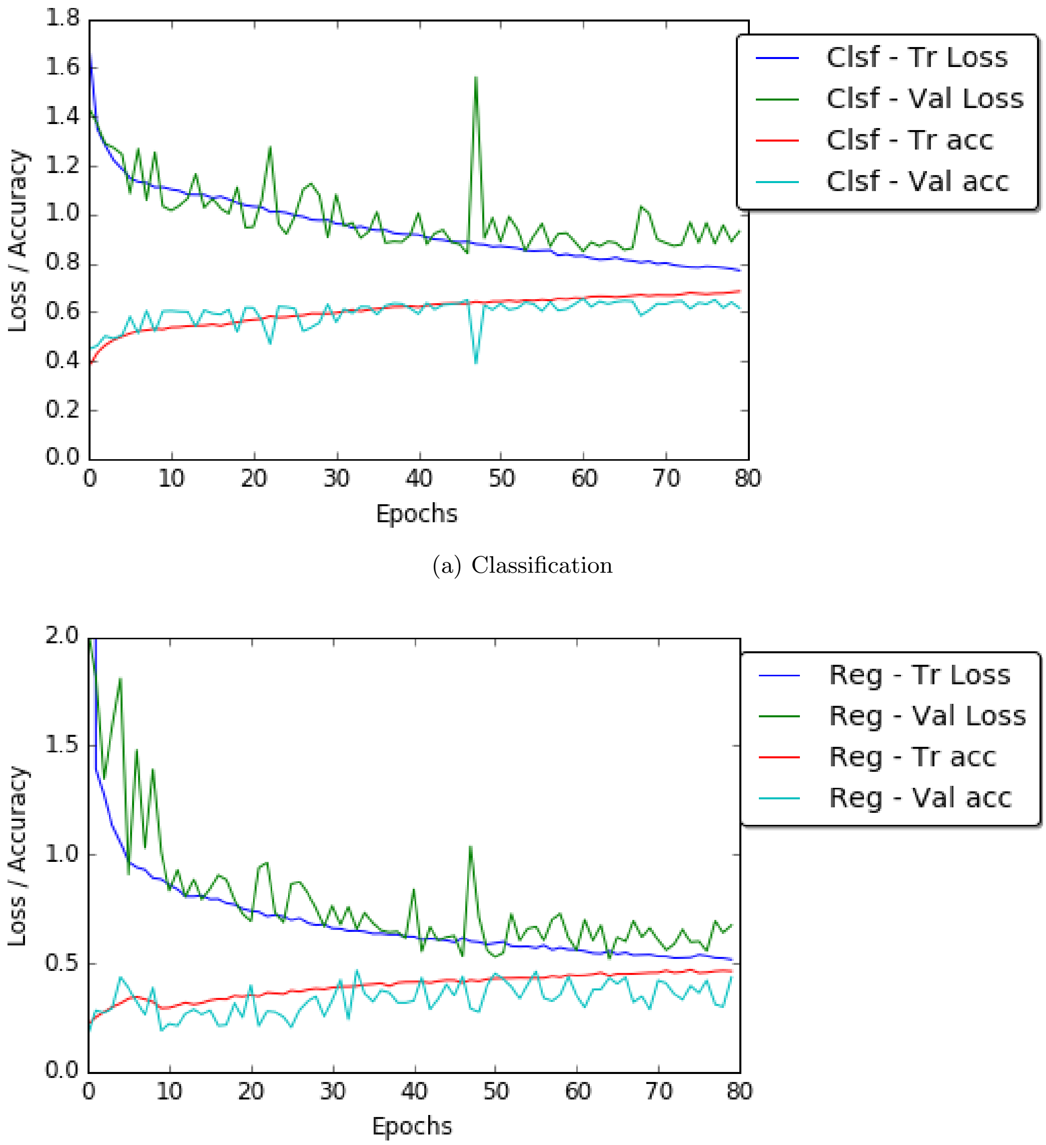
The input to the network are knee images of size 200×300. The knee images taken from the combined OAI-MOST dataset is used for training this network. The same train (70%) and test (30%) split are maintained from the previous experiments to make valid comparisons of the quantification results from the different methods. The right-left flip of the knee images is included to increase the training data and this doubles the training data. A validation split of 20% from the training data is used. This network is trained using the Adam optimise with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, as it gives faster convergence in comparison to the standard SGD.

Initial Results

Figure 5.8 shows the learning curves obtained whilst jointly training the CNN for classification and regression outputs. The learning curves show convergence in the validation and training losses with improvement in validation and classification accuracies. The jointly trained CNN with the initial configuration gives a classification accuracy of 60.8% with mean-squared error 0.795 for the classification outputs and 0.652 for the regression outputs. These results do not show improvement from the previous results. Previously, the network with the same configuration gave a classification accuracy of 61.8% and a mean-squared error 0.735 (Table 5.15) when trained to minimise only the classification loss. The same configuration after training to minimise only with the regression loss gave a mean squared error of 0.654 (Table 5.19). This configuration is optimal to minimise classification loss as it gave the highest classification accuracy (61.8%). However, this configuration is not optimal for regression as it gives a high mean-squared error (0.654). Next, the number of layers with learned weights and other hyper-parameters in this configuration are varied, to find a good architecture that will give improved results for both classification and regression outputs.

5.4.2 Tuning Hyper-parameters

The previous configuration does not improve the quantification performance and in particular the mean-squared error for regression output is high. Cascaded convolutional stages are included in the next configuration in an attempt to improve the regression outputs. The CNNs with cascaded convolutional stages gave best results for regression (Table 5.20) in the previous approach. Table 5.24



(a) Classification

(b) Regression

Figure 5.8: Learning curves for (a) classification and (b) regression in jointly trained CNN.

shows the network details. This network contains six layers of learned weights: five convolutional layers and a fully connected layer. The total free parameters in this network are ∼7.8 million. The other settings remain the same from the previous network and the same training procedure is followed.

This network gives a multi-class classification accuracy of 62.9% and the mean-squared error is 0.754 for the classification output and 0.583 for the

Table 5.24: Jointly trained network for classification and regression outputs.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	3	32×33×50
conv2-1	64	3×3	1	64×33×50
conv2-2	64	3×3	1	64×33×50
maxPool2	–	3×3	2	64×16×24
conv3-1	96	3×3	1	96×16×24
conv3-2	96	3×3	1	96×16×24
maxPool3	–	3×3	2	96×7×11
fc4	–	–	–	1024
fc5-Clsf	–	–	–	5
fc5-Reg	–	–	–	1

regression output. These results show improvement in the quantification performance in comparison to the previous results. Tuning the hyper parameters improves both the classification and the regression outcomes. Next, the depth of the architecture is increased and other related hyper parameters are tuned to investigate further improvement in the classification and regression outputs.

5.4.3 Best Performing Jointly Trained CNN

The best configuration (Table 5.25) is obtained after experimenting with different settings for jointly training a CNN for classification and regression outputs. This network has eight layers with learned weights: seven convolutional layers and a fully connected layer. This network has ∼2.9 million free parameters in total. This is a lightweight architecture with minimal parameters in comparison to the previous networks and the existing off-the-shelf CNNs. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. The fc5 layer is followed by two dense layers with softmax and linear activations for multi-class classification and regression outputs. To avoid overfitting, drop out with ratio 0.3 is included after the last fully connected (fc5) layer. Also, a L2 weight regularisation penalty of 0.01 is applied to all the convolutional and fully connected layers except the first two convolutional layers. This network is trained to minimise a weighted ratio of

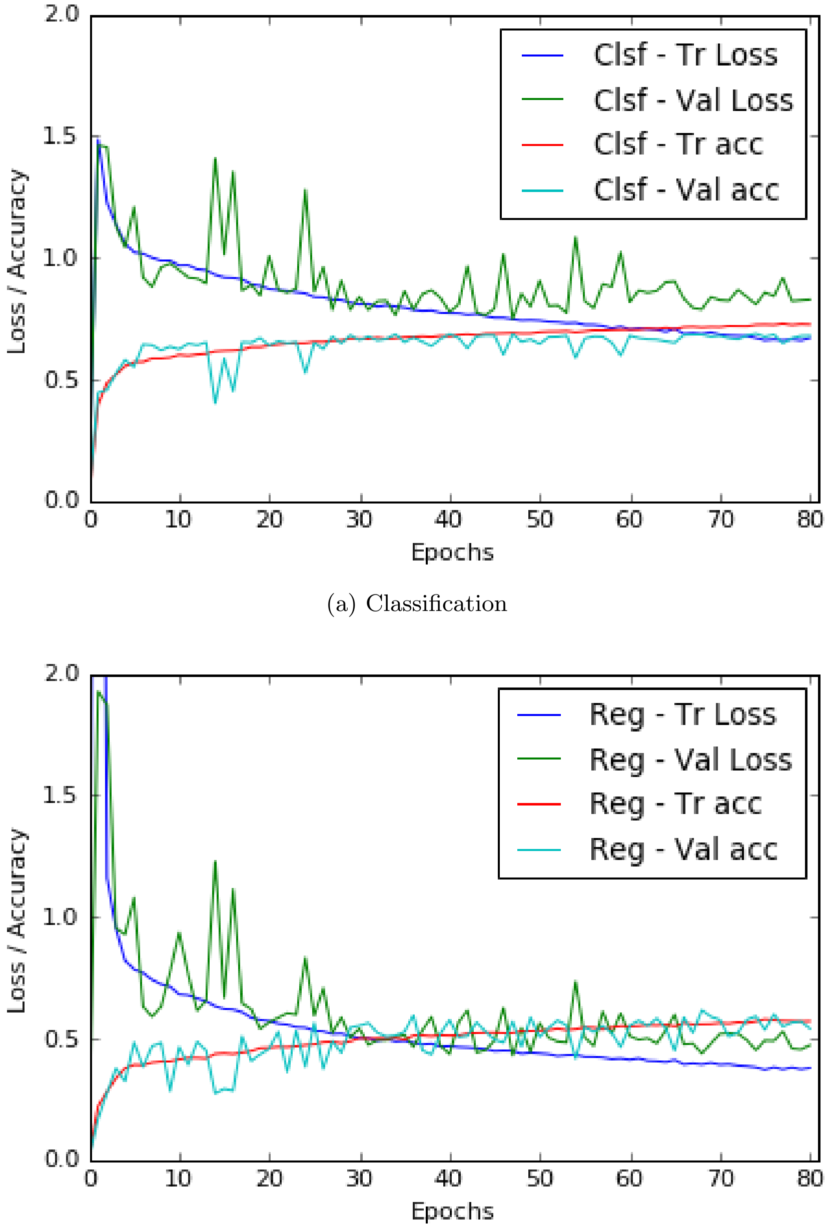
Table 5.25: Jointly trained network for classification and regression outputs.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2-1	64	3×3	1	64×49×74
conv2-2	64	3×3	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3-1	96	3×3	1	96×24×36
conv3-2	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4-1	128	3×3	1	128×11×17
conv4-2	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8
fc5	–	–	–	512
fc6-Clsf	–	–	–	5
fc6-Reg	–	–	–	1

two loss functions: categorical cross-entropy and mean-squared error. This network is trained using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$.

5.4.4 Jointly Trained CNN Results

Figure 5.9 shows the learning curves obtained whilst jointly training the CNN for classification and regression outputs. The learning curves show convergence to the minimum of the validation and training losses with improvement in validation and classification accuracies. The combined OAI-MOST dataset is used to compute these results. The same train-test split is maintained from the previous experiments. This jointly trained network gives a multi-class classification accuracy of 64.6% with a mean-squared error 0.685 for the classification outputs and 0.507 for the regression outputs. Table 5.26 shows the precision, recall, and F_1 score of this network. There is an improvement in the results: the classification accuracy increases to 64.6% from the initial configuration (60.8%), the mean-squared error for regression decreases to 0.507 from the initial configuration (0.652). Increasing the depth of the architecture by including more layers with



(a) Classification

(b) Regression

Figure 5.9: Learning curves for (a) classification and (b) regression in jointly trained CNN.

learned weights to the initial configuration and tuning the other hyper-parameters improves both the classification and regression results. Intuitively, providing a stronger error signal using both the classification and regression loss to the network allows to fit more parameters.

Table 5.26: Results of the best performing jointly trained CNN for classification and regression outputs.

Grade	Precision	Recall	F_1 Score
0	0.68	0.85	0.75
1	0.34	0.07	0.12
2	0.53	0.63	0.57
3	0.74	0.77	0.75
4	0.86	0.81	0.84
Mean	0.62	0.65	0.60

Table 5.27: Comparison of results from jointly trained CNN and individually trained CNNs for classification and regression results.

Method	Clsf-Accuracy	Clsf-MSE	Reg-MSE
CNN-Classification	61.8%	0.735	—
CNN-Regression	54.7%	—	0.574
Jointly trained CNN	**64.6%**	**0.685**	**0.507**

5.4.5 Results Comparison

The results of the jointly trained CNN are compared to the previous CNNs trained separately for classification and regression outputs. Table 5.27 shows the multi-class classification accuracy and mean-squared error of the jointly trained CNN and the separately trained CNNs for classification and regression outputs. There is an improvement in the classification accuracy and also the mean-squared error decreases for the joint training. These results show that the network jointly trained for classification and regression learns a better representation in comparison to the previous network trained separately for classification and regression outputs.

In summary, CNNs are trained from scratch to quantify knee OA severity using three approaches: classification, regression and jointly training for simultaneous classification and regression. From the results it is evident that the joint training outperforms both the individual training for classification and regression outputs. This supports the hypothesis that training a CNN for optimising a weighted ratio of two loss functions can improve the overall quantification of knee OA severity.

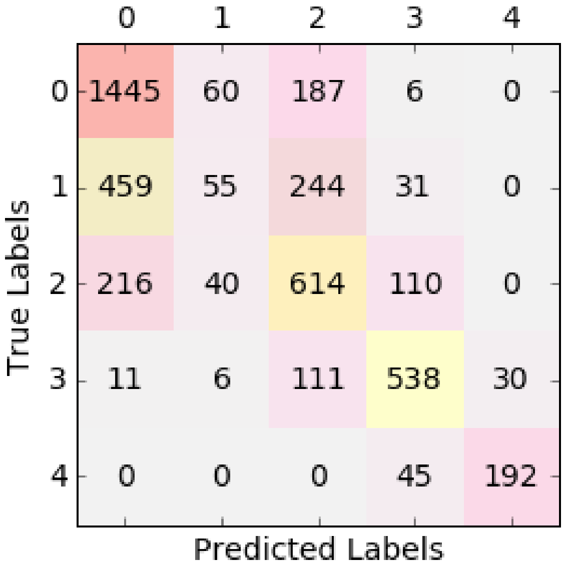


Figure 5.10: Confusion matrix for the multi-class classification using the jointly trained CNN.

5.4.6 Error analysis

A confusion matrix and the area under curve (AUC) after plotting the receiver operating characteristics are computed to perform an error analysis on the classification of the knee images by the jointly trained CNN. From the classification metrics (Table 5.26), the confusion matrix (Figure 5.10), and the receiver operating characteristic (ROC) curves (Figure 5.11), it is evident that classification of successive grades is challenging, and in particular classification metrics for grade 1 have low values in comparison to the other grades.

Figure 5.12 shows some examples of misclassification: grade 1 knee joints predicted as grade 0, 2, and 3. Figure 5.13 shows the misclassification of knee joints categorised as grade 0, 2 and 3 predicted as grade 1. These images show minimal variations in terms of joint space width and osteophytes formation, making them challenging to distinguish. Even the more serious misclassification in Figure 5.14, for instance grade 0 predicted as grade 3 and vice versa, do not show very distinguishable variations. Furthermore, when the knee X-ray images belonging to grade 0 and grade 1 severity are examined, it can be seen that there

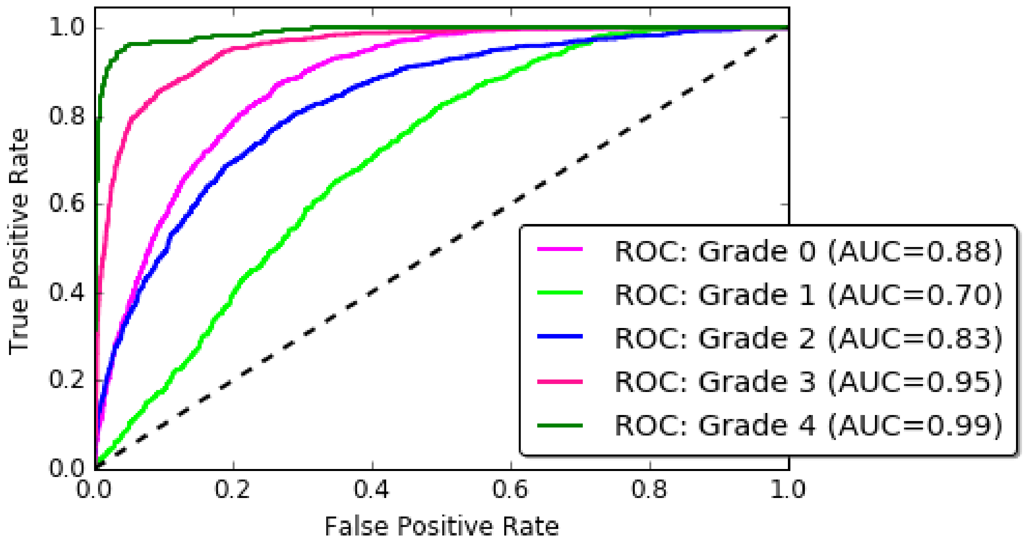


Figure 5.11: ROC for the multi-class classification using the jointly trained CNN.

are very subtle variations in terms of the joint space width and osteophytes formation. Even better representations are needed to capture these fine-grained variations and to distinguish coarse grades: grade 0 and grade 1 images.

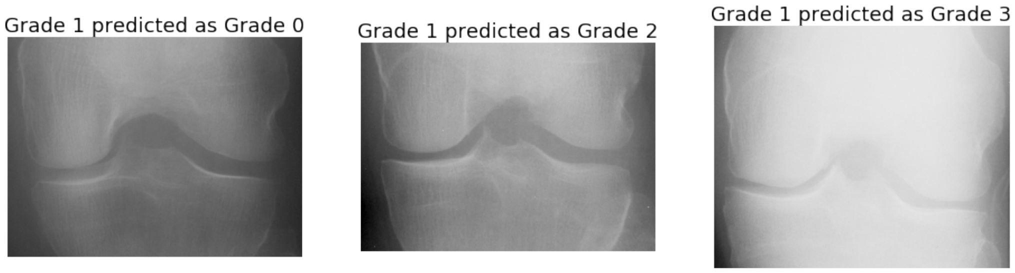

Grade 1 predicted as Grade 0
Grade 1 predicted as Grade 2
Grade 1 predicted as Grade 3

Figure 5.12: Mis-classifications: grade 1 joints predicted as grade 0, 2, and 3.

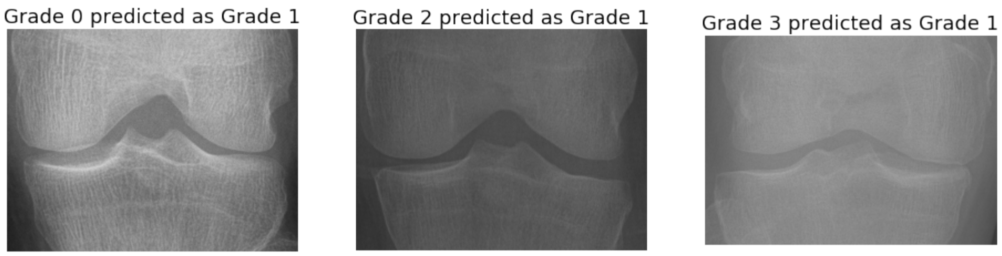

Grade 0 predicted as Grade 1
Grade 2 predicted as Grade 1
Grade 3 predicted as Grade 1

Figure 5.13: Misclassification: other grade knee joints predicted as grade 1.

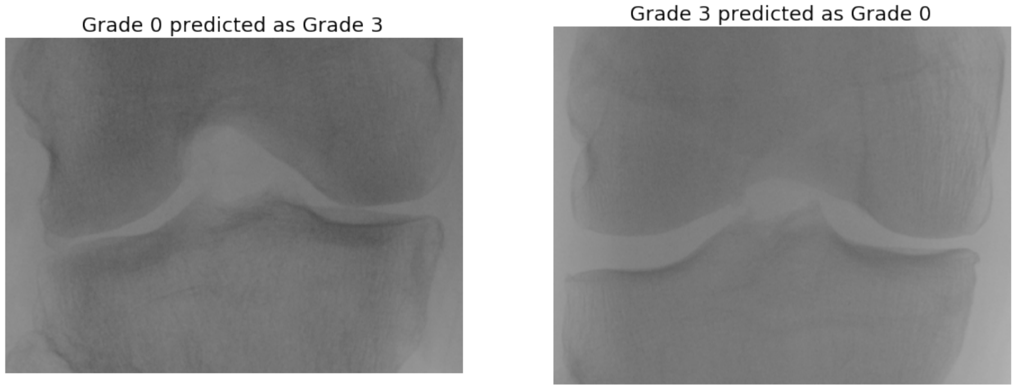


Figure 5.14: An instance of more severe misclassification: grade 0 and grade 3.

5.4.7 Discussion

Jointly training a CNN from scratch using the multi-objective convolutional approach improves the multi-class classification accuracy and minimises the mean-squared error. However, successive grade classification still remains a challenge. Even though the KL grades are widely used for assessing knee OA severity in clinical settings, there has been continued investigation and criticism over the use of KL grades as the individual categories are not equidistant from each other [5, 12, 13, 37, 38, 42]. This could be a reason for the low multi-class classification accuracy in the automatic quantification. Using OARSI readings instead of KL grades could possibly provide better results for automatic quantification as the knee OA features such as joint space narrowing, osteophytes formation, and sclerosis are separately graded. Moreover, when the knee X-ray images belonging to grade 0 and grade 1 severity are visually examined, it can be seen that there are very subtle variations in terms of the joint space width and osteophytes formation. To capture these variations and distinguish these coarse grades, for instance grade 0 versus grade 1, even better representations are required. Even medical experts do not always agree upon a particular KL grade e.g. either 0 or 1 attributed to the initial stage of knee OA [5, 13, 37, 38].

5.5 Ordinal Regression

Ordinal regression[1] is an intermediate task between multi-class classification and regression, sharing the properties of both. The outcomes or predictions in multi-class classification are discrete values and there is a meaningful order in the classes in regression. Ordinal regression is useful to classify patterns using a categorical scale which shows a natural order between the labels [35, 36]. The misclassification from a normal classifier are treated the same, that is no misclassification are worse than others [111]. Whereas, some misclassification in ordinal regression, for instance the misclassification on the extreme grades: grade 0 to grade 4 is treated worse than others. This implies that the distances between the classes need to be taken into account when training a classifier. When quantifying the stages of a physical disease, it is preferable to predict the stage as 'mild' or 'doubtful' than 'absent' when the true label is 'severe'. Ordinal regression models formalise this notion of order by ensuring that predictions farther from the true label incur a greater penalty than those closer to the true label [36]. The author believes that the KL grades prediction based on ordinal regression can further improve classification performance by reducing the margin of error (mean-squared error), considering the progressive nature of knee OA and the ground truth or labels for training a CNN i.e. the KL grades in an ordinal scale (0–4).

5.5.1 CNN Configuration for Ordinal Regression

For ordinal regression output, the last stage of the CNN (Table 5.25) that gave best results on the joint training for multi-class classification and regression is modified. The previous approach on the joint training used two dense layers with softmax and linear activations in parallel (Figure 5.7) for simultaneous multi-class classification and regression outputs. To train the CNN for ordinal regression, fixed weights ($[w_0, w_1, w_2, w_3, w_4] = [0, 1, 2, 3, 4]$) are applied to the outputs

[1] `https://statistics.laerd.com/spss-tutorials/ordinal-regression-using\ -spss-statistics.php`

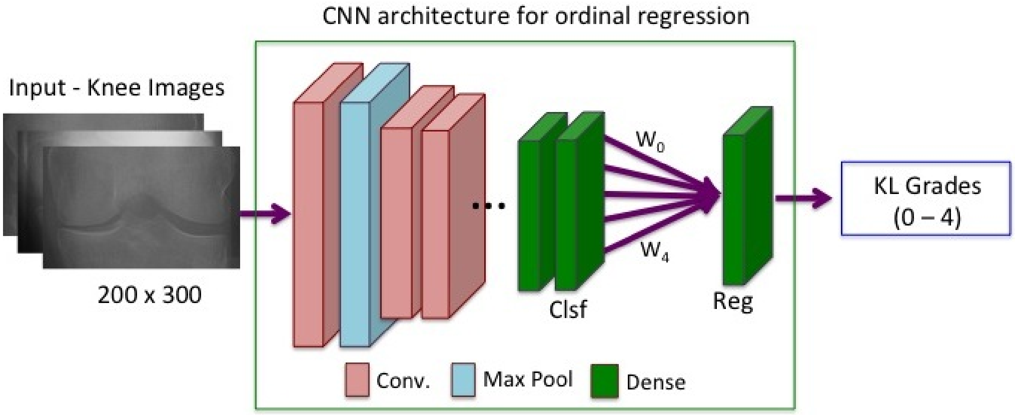


Figure 5.15: The CNN configuration for ordinal regression.

(probabilities) from the dense layer (Clsf) with softmax activations and back-propagate through a dense layer (Reg) with linear activations, optimising the mean-squared error loss function. The dense layer with softmax activations is treated as a hidden layer in this configuration. This is similar to the approach proposed by Beckham et al. [111] for ordinal classification. Figure 5.15 shows the CNN configuration for ordinal regression.

5.5.2 CNN Training

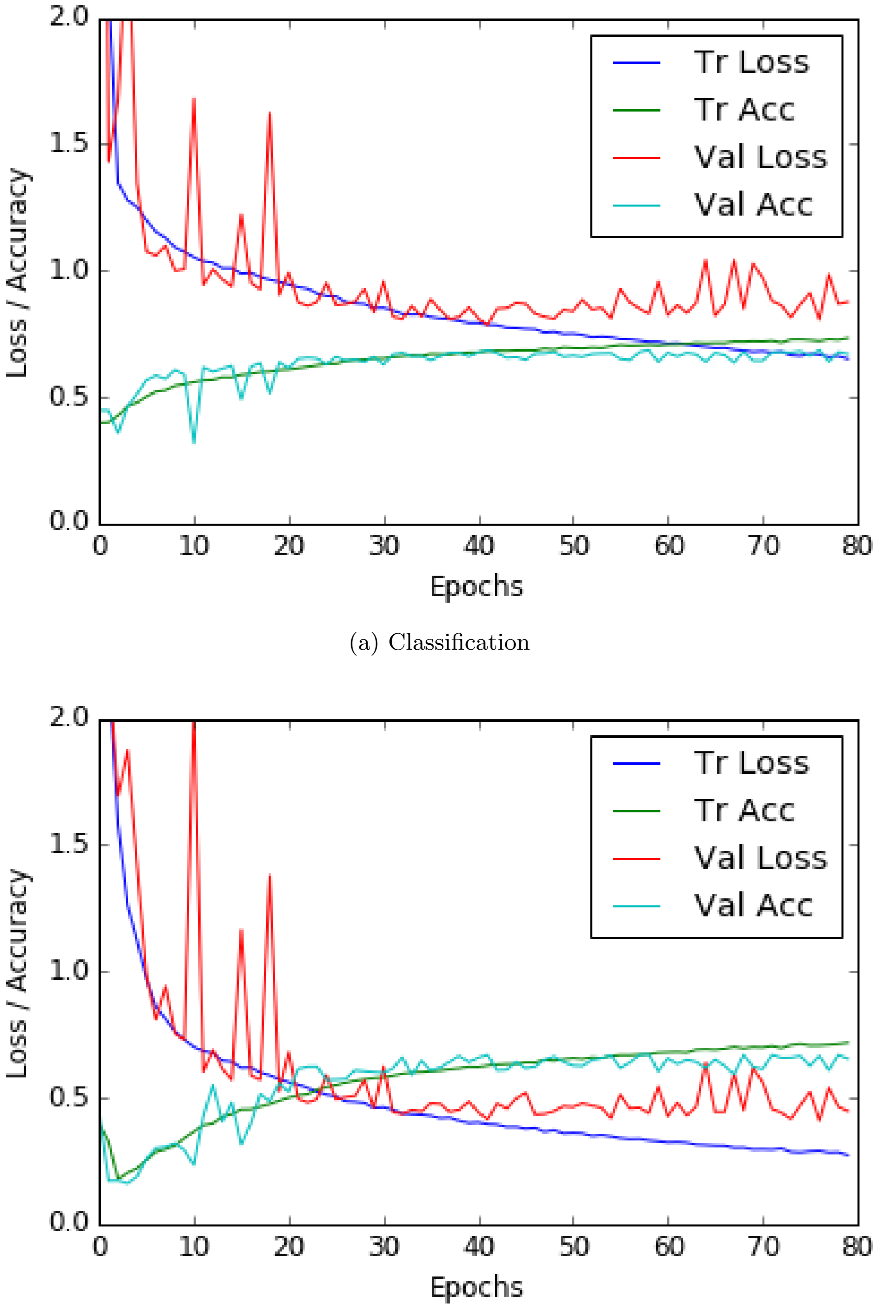
The CNN for ordinal regression (Table 5.28) is based on a lightweight architecture with ∼2.9 million free parameters in total and it contains eight layers with learned weights: seven convolutional layers and a fully connected layer. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. To avoid overfitting, drop out with ratio 0.3 is applied after the last fully connected (fc5) layer. In addition to this, a L2 weight regularisation penalty of 0.01 is applied to all the convolutional and fully connected layers except the first two convolutional layers. The fc5 layer is followed by two dense layers with softmax (fc6-Clsf) and linear activations (fc7-Reg). The output of the softmax (fc6-Clsf) layer is multiplied (dot product) with fixed weights ([0,1,2,3,4]) and given as input to the last dense

Table 5.28: CNN architecture for ordinal regression

Layer	Kernels	Kernel Size	Strides	Output shape
input	–	–	–	1×200×300
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2-1	64	3×3	1	64×49×74
conv2-2	64	3×3	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3-1	96	3×3	1	96×24×36
conv3-2	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4-1	128	3×3	1	128×11×17
conv4-2	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8
fc5	–	–	–	512
fc6-Clsf	–	–	–	5
input-weights	–	–	–	5
merge-product	–	–	–	1
fc7-Reg	–	–	–	1

layer (fc7-Reg). This network is trained to minimise two loss functions: categorical cross-entropy and mean-squared error with equal weights. The dense layers (fc7-Reg) and (fc6-Clsf) provides the ordinal regression and multi-class classification outputs. The network is trained for 80 epochs with a batch size 32, using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. The same training, validation, and test data are used from the joint training to make valid comparison of the results.

The CNN configuration in Table 5.28 gives the best results for ordinal regression and this configuration is similar to the network for joint training (Table 5.25 except the arrangement of the last two dense layers. Tuning the hyper-parameters of this CNN by increasing the number of layers with learned weights does not improve the quantification performance. Therefore, this CNN configuration is selected as the final network for ordinal regression.



(a) Classification

(b) Ordinal regression

Figure 5.16: Learning curves for (a) classification and (b) ordinal regression .

5.5.3 Results

The learning curves (Figure 5.16) obtained whilst training the CNN for ordinal regression shows convergence of the validation and training losses with

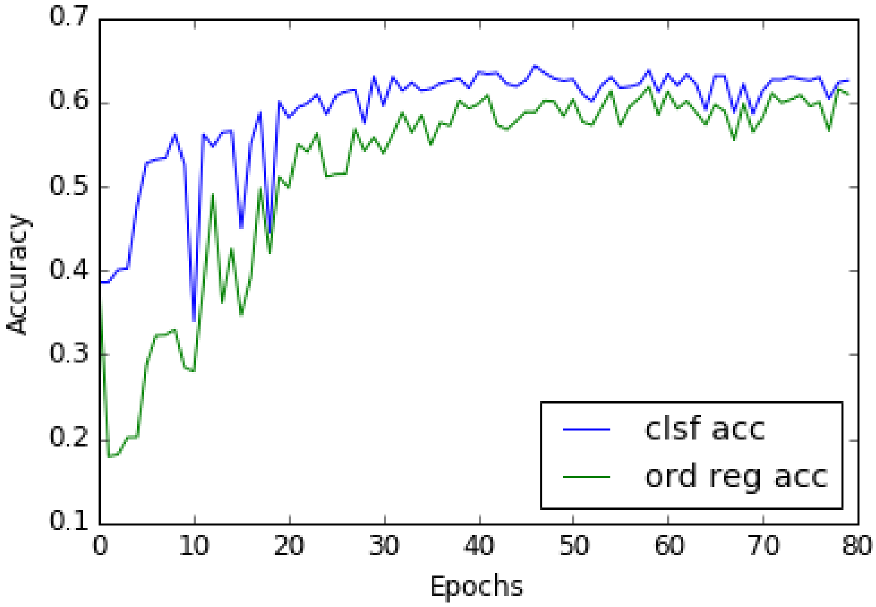


Figure 5.17: The CNN configuration for ordinal regression.

improvement in validation and classification accuracies. Figure 5.17 shows the classification accuracy of the trained CNN model after every epoch on the test data for the ordinal regression and classification output. After 40 epochs of training, there is no significant improvement in the classification accuracies. There is a slight decrease in the accuracy of the ordinal regression in comparison to the classification. This is likely due to the rounding of the output.

After training, the CNN gives a classification accuracy of 64.3% on the test data. In the previous method on jointly training a CNN for classification and regression (Section 5.4.4), a multi-class classification accuracy of 64.6% (Table 5.26) is achieved. As the same CNN configuration is used except the last stage (Figure 5.15) and other settings are retained, the classification performance remains almost the same for the CNN trained for ordinal regression in comparison to the jointly trained CNN.

The classification metrics for the ordinal regression output are computed by rounding the predictions to integer values (0, 1, 2, 3, or 4). After rounding, the classification accuracy for the ordinal regression output is 61.8% with mean-squared error 0.504 on the test data. The classification metrics for the

Table 5.29: Comparison of classification metrics from regression and ordinal regression.

Grades	Regression			Ordinal Regression		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.78	0.58	0.66	0.71	0.79	0.75
1	0.29	0.55	0.38	0.31	0.33	0.32
2	0.51	0.50	0.50	0.59	0.44	0.50
3	0.65	0.53	0.58	0.74	0.73	0.73
4	0.63	0.33	0.43	0.81	0.76	0.78
Mean	0.60	0.53	0.55	0.62	0.62	0.61

regression output from the jointly trained CNN gives a classification accuracy 53.3% with mean squared error 0.595 on the test data. There is an improvement in the classification accuracy and mean-squared error for ordinal regression in comparison to the previous regression results. Table 5.29 shows the precision, recall, and F_1 score for regression and ordinal regression. From the results it is evident that ordinal regression is out performing regression for quantifying knee OA images in a continuous scale.

5.6 Results Comparison

Four approaches are investigated to automatically quantify knee OA severity. Table 5.30 shows the multi-class classification accuracy and mean-squared error for the four approaches: 1) fine-tuning off-the-shelf CNN (BVLC CaffeNet), 2) training CNNs from scratch individually for classification and regression, 3) jointly training a CNN based on multi-objective convolutional learning for classification and regression, and 4) training a CNN for ordinal regression. These results are compared to WNDCHRM, that gave the previous best results for automatically classifying knee OA radiographs. The results show that jointly trained CNN gives best results for multi-class classification. The ordinal regression outperforms all the other methods for quantifying knee OA images in a continuous scale with low mean-squared error.

Table 5.30: Comparison of classification, regression and ordinal regression results.

Method	Classification Accuracy	Mean-Squared Error
WNDCHRM	34.8%	2.112
Fine-Tuned BVLC CaffeNet	57.6%	0.836
CNN-Classification	61.8%	0.735
CNN-Regression	54.7%	0.574
Jointly trained CNN	**64.6%**	0.507
Ordinal Regression	64.3%	**0.480**

5.7 Chapter Summary

Four approaches are presented in this chapter to automatically assess knee OA severity using CNNs. First, the existing pre-trained CNNs are investigated for classifying knee images based on KL grades. Two methods are used in this approach: using the pre-trained CNNs for fixed feature extraction, and fine-tuning the pre-trained CNNs using the transfer learning approach. The predictions or outputs from these methods are ordinal KL grades (0,1,2,3 or 4). Furthermore, the author argued that quantifying knee OA severity in a continuous scale (0–4) is more appropriate as the OA degradation is progressive in nature, not discrete. Regression is used to quantify the knee OA severity on a continuous scale. The classification and regression results from the proposed methods in this chapter outperform the previous best results achieved by WNDCHRM, which uses many hand-crafted features with a variation of k-nearest neighbour classifier for classifying knee OA radiographs.

Second, CNNs are trained from scratch for classification and regression. The objective was to further improve the quantification results. As the training data is relatively scarce, a lightweight architectures with fewer (∼4 to ∼5 million) free parameters are used in the CNNs. The fully trained CNN for classification achieved high classification accuracy in comparison to the pre-trained CNNs. However, the fully trained CNN for regression did not achieve high-performing results as no ground truth of KL grades was available on a continuous scale. Therefore, the discrete KL grades are used to train the CNNs for regression.

Third, CNNs are fully trained using multi-objective learning for simultaneous classification and regression. The intuition behind this is optimising a CNN with two loss functions provide a stronger error signal and it is a step to improve the overall quantification, considering both classification and regression results. The jointly trained CNN achieved better quantification results with a high classification accuracy in comparison to the previous methods.

As the last approach, CNNs are fully trained for ordinal regression using a softmax dense layer as the hidden layer. This approach achieved low mean-squared error and outperformed other methods to quantify knee OA severity in a continuous scale. The added benefit of this method is to provide simultaneous multi-class classification output.

In summary, a progressive improvement is achieved in the quantification performance with an increase in classification accuracy and other performance metrics in the four approaches to automatically quantify knee OA severity. To conclude this chapter, an error analysis is presented that discusses the possible reasons for the misclassification from the jointly trained CNN. The variations in the X-ray imaging protocols and discrepancies in the KL grades scoring needs to be taken into account when analysing the misclassification.

There are criticisms in the literature over the use of KL grades for knee OA assessment [5, 12, 13, 37, 38] and also there are claims stating that OARSI readings are more accurate than KL grades [39, 42]. Therefore, OARSI readings are investigated to quantify the individual knee OA features such as joint space narrowing and osteophytes formation for knee OA assessment. CNNs are trained for this and the results are analysed in the next chapter.

Chapter 6

Automatic Quantification of Knee OA Diagnostic Features

6.1 Introduction

The previous chapter focused on training CNNs to quantify knee OA severity on an ordinal scale using classification and on a continuous scale using regression over the discrete KL grades. In this chapter, the focus is on the approaches to quantify distinct knee OA features such as joint space narrowing (JSN) and osteophytes based on the OARSI readings in an attempt to improve the overall quantification. First, CNNs are trained to quantify JSN and osteophytes individually. Next, CNNs are trained using multi-objective convolutional learning to jointly quantify the distinct knee OA features and KL grades. To conclude this chapter, an automatic knee OA diagnostic system is proposed by combining the automatic localisation pipeline that was developed in Chapter 4 and the quantification pipeline that is developed in this chapter.

There are claims in the literature that KL grades are not sufficiently accurate and reliable for radiographic classification of knee OA [5, 12, 13, 37, 38]. Moreover, some studies in the literature claim that the OARSI readings that grade the distinct knee OA radiographic features such as JSN, osteophytes, sclerosis, and attrition are more accurate and reliable for assessing knee OA severity [39, 42]. Therefore, the next approach investigates the assessment of distinct knee OA radiographic features using OARSI readings as the ground truth.

First, CNNs are trained from scratch to classify knee OA radiographs using the OARSI gradings for JSN and osteophytes as ground truth. OARSI grades the individual knee OA features: JSN and osteophytes in an ordinal scale $(0 - 3)$, where 0-normal, 1-mild, 2-moderate, 3-severe. CNNs are also jointly trained to quantify knee OA severity based on JSN, osteophytes and KL grades using multi-objective convolutional learning in an attempt to further improve the overall quantification.

Next, a knee OA diagnostic system is developed combining the two pipelines: the automatic localisation using the FCN to localise the knee joints from radiographs and the quantification using the jointly trained CNN for simultaneous classification and regression of the localised knee joints. The performance of the proposed system is compared to the gold standard: the KL and OARSI readings of knee radiographs from the OAI. To test the reliability of the KL and OARSI readings, the evaluators in OARSI have used simple kappa[1] and weighted kappa coefficients to evaluate the agreement between the readings when the variables were assigned more than two ordinal categories.

The remainder of this chapter is organised as follows: Section 6.2 introduces the dataset and the ground truth used for the experiments in this chapter. Section 6.3 and Section 6.4 presents the deep learning approaches to quantify distinct knee OA features: JSN and osteophytes. Section 6.5 presents the multi-objective convolutional learning approach to simultaneously quantify the associated knee OA features with KL grades. Section 6.6 compares and analyses the results of the CNNs trained to quantify distinct knee OA features individually and jointly. Section 6.7 compares the performance of the proposed system to the existing gold standard for both multi-class and binary classifications. Section 6.8 proposes a practical knee OA diagnostic system. Section 6.7 summarises this chapter and presents the conclusions.

[1] Cohen's kappa coefficient measures the inter-rater agreement for qualitative or categorical classifications.

6.2 Dataset and Ground Truth

The OAI and the MOST datasets are used for the experiments in this chapter. These datasets contain the radiographic KL scores and the OARSI radiographic assessment readings for the knee OA clinical features such as JSN, osteophytes formation, subchondral sclerosis, chondrocalcinosis, cysts, and attrition. All these features are integrated in composite scoring systems like Kellgren & Lawrence (KL) grading system [1, 3, 5, 112].

Figure 6.1 shows a knee radiograph with healthy cartilage, and JSN between the femur and tibia bones due to cartilage loss. The knee radiographs are vertically split into two halves to assess the distinct knee OA features separately in the lateral and medial compartments. Figure 6.1 shows the lateral and medial compartments in a knee radiograph. The knee OA features: osteophytes formation, subchondral sclerosis, and cysts, are separately assessed for the femur and tibia bones on both the lateral and the medial compartments. Cysts and chondrocalcinosis are scored in a binary scale (0 & 1) and the rest of the features are graded in an ordinal scale (0–3).

The OAI and the MOST datasets contain 4,746 and 3,026 participants in total. The OARSI readings are not available for many participants. In total 10,861 ($\sim$70%) knee joints are selected from the combined OAI-MOST dataset based on the availability of the assessments for KL grades, JSN, and osteophytes. Table 6.1 shows the grade-wise distribution of the knee joints as per the KL grades and the knee OA features: JSN and osteophytes for the radiographs in the combined OAI-MOST dataset. The distribution shows that there is imbalance in the datasets and the number of samples in the severe grades (≥ 2) are low. The number of samples with OARSI assessment readings for sclerosis, cysts, chondrocalcinosis and attrition is even lower ($< 35\%$). Therefore, the experiments are restricted to the knee OA clinical features: JSN and osteophytes.

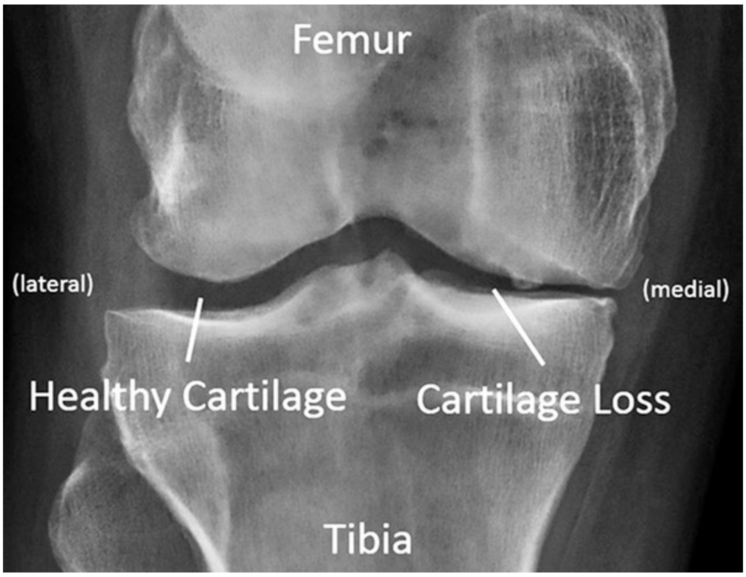


Figure 6.1: A knee radiograph showing healthy cartilage in the lateral compartment and joint space narrowing (JSN) due to cartilage loss in the medial compartment.

Source: `http://lermagazine.com/article/gait-retraining-improves-symptoms-of-knee-oa`

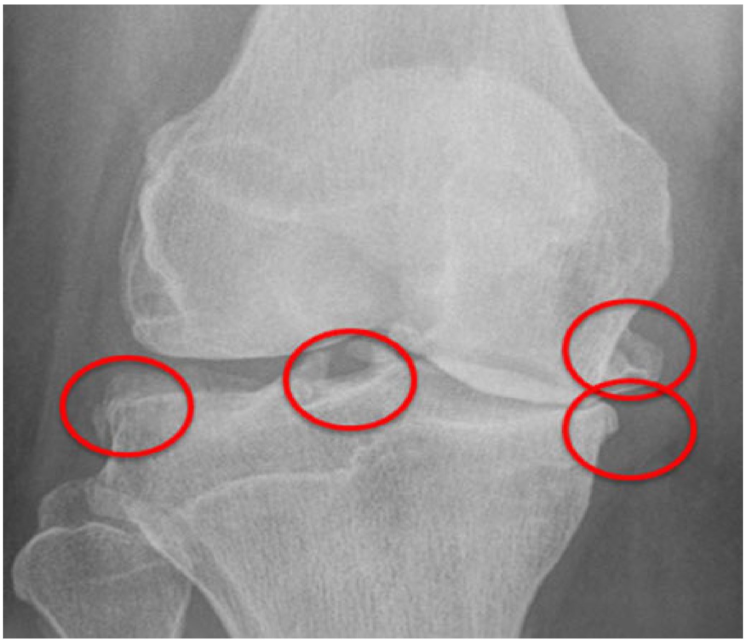


Figure 6.2: A few instances of visible femoral and tibial osteophytes in the lateral and medial regions.

Source: `http://www.melbournekneeortho.com.au/case-studies/knee-arthritis`

Table 6.1: Grade-wise distribution of the knee joints as per KL grades, JSN, and osteophytes for the radiographs in the combined OAI-MOST dataset.

Grades	KL scores	JSN		Osteophytes			
				Femur		Tibia	
		Lat	Med	Lat	Med	Lat	Med
0	3,146	9,695	6,097	7,589	6,781	7,507	5,263
1	1,746	555	2,581	1,784	1,678	2,265	4,111
2	3,227	434	1,707	802	975	589	1,000
3	2,091	177	476	686	1,427	500	487
4	651	–	–	–	–	–	–

6.3 Training CNNs to Quantify JSN

The previous approach (in Chapter 5) focused on training CNNs from scratch to quantify knee OA severity based on KL grades. Some studies in the literature claim that the measurement of the radiographic joint space width is the most accepted and a suitable method for assessing the progression of knee OA [12,39,42]. As it has been shown to be sensitive to small changes, joint space narrowing (JSN) remains the primary outcome by which disease modifying OA drug trials test drug efficacy. Therefore, this approach focuses on training CNNs from scratch to quantify JSN from knee radiographs. JSN is scored separately for the medial and the lateral compartments in a knee radiograph. Figure 6.1 shows the medial and the lateral compartments of a knee joint in a radiograph. CNNs are trained individually to quantify lateral JSN and medial JSN.

6.3.1 CNN Configuration and Training Process

A similar architecture is used from the previous approach (Section 5.3.6) to quantify JSN in the lateral compartment of the knee joints. Table 6.2 shows the CNN architecture in detail. This CNN gave the best results for classifying knee images based on the KL grades. All the previous settings are retained except the last fully connected layer (fc6) is replaced with a layer with 4 outputs as the ground truth (JSN lateral) contains 4 categories. Drop out after conv4 and fc5

Table 6.2: CNN for classifying knee images based on lateral JSN.

Layer	Kernel	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2	96	5×5	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3	128	3×3	1	128×24×36
maxPool3	–	3×3	2	128×11×17
conv4	256	3×3	1	256×11×17
maxPool4	–	3×3	2	256×5×8
fc5	–	–	–	1024
fc6	–	–	–	4

layers, and L2 weight regularisation in all convolutional and fully connected layers are used. The network is trained to minimise categorical cross-entropy using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}^{-8}$. The input to the network are localised knee images of size 200×300. Table 6.1 shows the number of images available in the combined OAI and MOST datasets. The dataset is split into training (70%) and test (30%) sets. The validation (20%) set is taken from the training samples.

6.3.2 Classification Results for Lateral JSN

After training, this network gives a multi-class classification accuracy of 90.8% on the test data and the mean-squared error is 0.154. Table 6.3 shows the precision, recall, and F_1 score of this network for classification based on lateral JSN. The results show high multi-class classification accuracy (90.8%) and high mean F_1 score (0.9). The precision, recall and F_1 score of lateral JSN belonging to grade 3 is 0. This is due to scarce training samples in grade 3. Nevertheless, the mean F_1 score and classification accuracy are high due to the imbalance in the datasets. The grade wise distribution of the images for lateral JSN is grade 0 - 9,695; grade 1 - 555; grade 2 - 434; and grade 3 - 177. The number of training samples in total are very high for grade 0 and very low for grade 3.

Table 6.3: Results for classifying knee images based on lateral JSN.

Grade	Precision	Recall	F_1 Score
0	0.96	0.98	0.97
1	0.53	0.22	0.31
2	0.38	0.74	0.50
3	0.0	0.0	0.0
Mean	0.90	0.91	0.90

The balanced accuracy is calculated to evaluate this classification given this unbalanced dataset. The balanced accuracy is the average of per-class classification accuracy. The balanced accuracy for the multi-class classification of this network is 48.6%. The low balanced accuracy again shows that this classification is biased due to the imbalance in the datasets. To overcome this limitation and in an attempt to improve the classification performance, multi-objective convolutional learning is investigated in Section 6.5.

6.3.3 Classification Results for Medial JSN

Classifying knee images based on lateral JSN gives low balanced accuracy as there is insufficient training samples belonging to high grades. A CNN is trained to classify knee images based on medial JSN. The same network configuration (Table 6.2) is used from the previous settings. The grade-wise distribution of the knee images as per medial JSN in the combined OAI-MOST dataset is grade 0 - 6,097; grade 1 - 2,581; grade 2 - 1,707; and grade 3 - 476. This distribution shows that there are relatively more training samples in higher grades in comparison to the previous distribution of lateral JSN knee images. The dataset is split into training (70%) and test (30%) sets. The validation (20%) set is taken from the training samples.

This network (Table 6.2) gives a multi-class classification accuracy of 76.5% on the test data. The balanced accuracy is 72.3%. The classification results show high multi-class classification accuracy and high balanced accuracy. Table 6.4 shows the precision, recall, and F_1 score for the classification based on medial JSN. There is

Table 6.4: Results for classifying knee images based on medial JSN.

Grade	Precision	Recall	F_1 Score
0	0.82	0.91	0.86
1	0.62	0.37	0.46
2	0.72	0.81	0.76
3	0.63	0.89	0.74
Mean	0.75	0.77	0.75

improvement in the classification results of medial JSN (Table 6.4) in comparison to the previous results (Table 6.3) for classification of lateral JSN knee joints with the same network configuration. This is due to the availability of sufficient training data.

In summary, CNNs are trained with the same configuration to classify knee images based on JSN in lateral and medial compartments. As there is sufficient training samples for high grades of medial JSN, the network gives high balanced accuracy. On the other hand, the classification performance is low for lateral JSN due to insufficient training data in high grades. Next, classification of knee images is investigated to automatically quantify femoral and tibial osteophytes.

6.4 Training CNNs to Quantify Osteophytes

Osteophytes or bone spurs is a characteristic feature that defines the presence of radiographic knee OA degradation and increases the risk of structural knee OA progression [5,112]. The OARSI scoring system classifies the presence of osteophytes according to the four regions of the knee joint as lateral femur, medial femur, lateral tibia and medial tibia. Figure 6.2 shows a few instances of osteophytes presence in these regions. The OARSI scores for the osteophytes is an ordinal scale (0–3) like the JSN. CNNs are trained to quantify the four distinct osteophytes individually.

Table 6.5: Results for classifying knee images based on lateral femoral osteophytes.

Grade	Precision	Recall	F_1 Score
0	0.82	0.90	0.86
1	0.31	0.20	0.24
2	0.25	0.07	0.11
3	0.37	0.66	0.48
Mean	0.66	0.70	0.67

6.4.1 Quantifying Femoral Osteophytes

CNNs are trained to classify knee images based on femoral osteophytes in lateral and medial compartments. The same architecture (Table 6.2) is used from the previous approach as this network gave the best results for classification of KL grades and JSN medial. The grade-wise distribution of the knee images in the combined OAI-MOST datasets according to lateral femoral osteophytes is grade 0 - 7,589; grade 1 - 1,784; grade 2 - 802; and grade 3 - 686. The grade-wise distribution for medial femoral osteophytes is grade 0 - 6,781; grade 1 - 1,678; grade 2 - 975; and grade 3 - 1,427. The dataset is split into training (70%) and test (30%) sets. The validation (20%) is taken from the training set.

6.4.2 Classification Results for Lateral Femoral Osteophytes

The multi-class classification accuracy of this network is 70.3% for classifying femoral osteophytes in the lateral compartment of knee. The balanced classification accuracy is 45.7%. The results show high multi-class classification accuracy and low balanced accuracy. Table 6.5 shows the precision, recall, and F_1 score of this network for classification of femoral osteophytes in the lateral compartment. The precision, recall, and F_1 score are low for grade 1 and grade 2 classifications. The network is not learning effective representations to pick up small variations in the presence of femoral osteophytes in lateral compartments belonging to consecutive grades. Next, the classification of knee images is investigated based on femoral osteophytes in the medial compartment.

Table 6.6: Results for classifying knee images based on medial femoral osteophytes.

Grade	Precision	Recall	F_1 Score
0	0.80	0.92	0.86
1	0.35	0.25	0.29
2	0.0	0.0	0.0
3	0.56	0.73	0.63
Mean	0.63	0.71	0.66

6.4.3 Classification Results for Medial Femoral Osteophytes

The same network (Table 6.2) trained for classifying femoral osteophytes in the medial compartment gives a multi-class classification accuracy of 70.8%. The balanced classification accuracy is 47.4%. Table 6.5 shows the precision, recall, and F_1 score of this network for classification of femoral osteophytes in the medial compartment. These results are similar to the results for classification of femoral osteophytes in the medial compartment (Table 6.5). These results again show that the network is not learning effective representations to classify consecutive grades of osteophytes severity.

6.4.4 Discussion

CNNs are trained with the same configuration (Table 6.3) keeping the previous settings to classify tibial osteophytes in the lateral and the medial compartments of the knees. The multi-class classification accuracy is 70.2% and the balanced accuracy is 45.3% for the lateral tibial osteophytes. The network gives a classification accuracy of 57.3% and a balanced accuracy of 46% for medial tibial osteophytes.

The overall results for classification of femoral and tibial osteophytes give low balanced accuracy. This is due to the imbalance in the datasets and the fact that there is insufficient training samples for high grade osteophytes. The joint classification of JSN, osteophytes and KL grades is investigated using multi-objective convolutional learning, on the assumption that multi-objective optimisation will improve the overall quantification results.

6.5 Jointly Training a CNN to quantify JSN, Osteophytes and KL grades

In the previous chapter, CNNs were jointly trained to minimise a weighted ratio of two loss functions for simultaneous classification and regression of KL grades following the multi-objective convolutional approach. This joint training improved both the classification and regression results. Motivated by this, CNNs are jointly trained to classify knee images based on KL grades along with the distinct knee OA features: JSN and osteophytes. In addition to this, the other reasons for investigating multi-objective convolutional learning for this joint classification are as follows.

- **Multi-objective optimisation:** As discussed earlier, the KL grades is a composite scoring that takes into account the distinct knee OA features such as JSN, osteophytes, sclerosis, cysts and attrition [1,3,5,112]. Training a CNN with multiple ground truth annotations such as JSN and osteophytes, along with KL grades belonging to the same domain can provide more information to the CNN for optimising multiple closely related objectives. The author believes that jointly training a CNN will improve the quantification of KL grades as well as the JSN and osteophytes.

- **Addressing the multi-label imbalance problem:** The OARSI readings for the knee OA features and KL grades are used as the ground truth to jointly train the CNNs. The number of training samples in the severe grades of the knee OA features such as sclerosis, cysts, chondrocalcinosis and attrition are very low. For this reason the investigations are restricted to the knee OA features: JSN and osteophytes. Still there is an imbalance in the datasets for JSN and osteophytes. The author believes that the jointly trained network with a loss function that mixes multiple objectives with domain adaptive weighting of the propagated loss can address the multi-label imbalance problem to a certain extent. There are studies in the literature that report jointly trained CNNs based on the multi-objective convolutional learning as a solution for the multi-label data imbalance problem [106, 110].

6.5.1 Initial Configuration

A similar architecture is used from the previous approach (Section 5.4.3) as an initial configuration to jointly train a CNN for simultaneous quantification of JSN, osteophytes and KL grades. The configuration in Table 6.7 gave the best results for simultaneous classification and regression of KL grades on a validation set. The fully connected layer (fc5) is followed by seven independent softmax layers to give the multi-class classification outputs for KL grades (fc6-KL), lateral JSN (fc6-LatJSN) and medial JSN (fc6-MedJSN), femoral osteophytes in lateral (fc6-FemLatOst) and medial compartments (fc6-FemMedOst), and tibial osteophytes in lateral (fc6-TibLatOst) and medial (fc6-TibMedOst) compartments. Figure 6.3 shows the configuration details of this network. To avoid over-fitting, a drop out layer with drop out ratio 0.3 is included after the last fully connected (fc5), and a L2-norm weight regularisation penalty of 0.01 is applied to all the convolutional and fully connected layers except the first two convolutional layers. This network is trained to minimise the cumulative sum of categorical cross-entropy for the seven multi-class classifications with equal weights. This network is trained using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, β_1 = 0.9, β_2 = 0.999, ϵ = 1e^{-8}. The input to the network are knee images of size 200×300 taken from the OAI and the MOST datasets. Table 6.1 shows the grade-wise distribution of the knee images in the combined OAI and MOST datasets. The dataset is split into training (70%) and test (30%) sets. 20% of the training samples are used for validation whilst training the network.

Table 6.7: Initial configuration of the jointly trained network for classification of knee images based on KL grades, JSN, and osteophytes.

Layer	Kernel	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2-1	64	3×3	1	64×49×74
conv2-2	64	3×3	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3-1	96	3×3	1	96×24×36
conv3-2	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4-1	128	3×3	1	128×11×17
conv4-2	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8
fc5	–	–	–	512
fc6-KL	–	–	–	5
fc6-LatJSN	–	–	–	4
fc6-MedJSN	–	–	–	4
fc6-FemLatOst	–	–	–	4
fc6-FemMedOst	–	–	–	4
fc6-TibLatOst	–	–	–	4
fc6-TibMedOst	–	–	–	4

6.5.2 Results from Initial Configuration

Figure 6.4 shows the learning curves for KL grades classification whilst jointly training this CNN (Table 6.7). The learning curves show gradual increase in the validation and training accuracies with decrease in the validation and training losses. Table 6.8 shows the classification results for the joint training: accuracy, balanced accuracy, mean average precision, mean recall, and mean F_1 score on the test set. The balanced accuracy is calculated as there is high imbalance in the datasets. The results show small improvement in the KL grade classification accuracy (64.9%) in comparison to the previous results (Table 5.27 on Page 125) that gave classification accuracy (64.6%) for jointly training the CNN for classification and regression losses. The classification accuracy, mean precision, recall, and F_1 score for multi-class classification of JSN and osteophytes are high, but the balanced classification accuracy is still low. This is due to the dataset

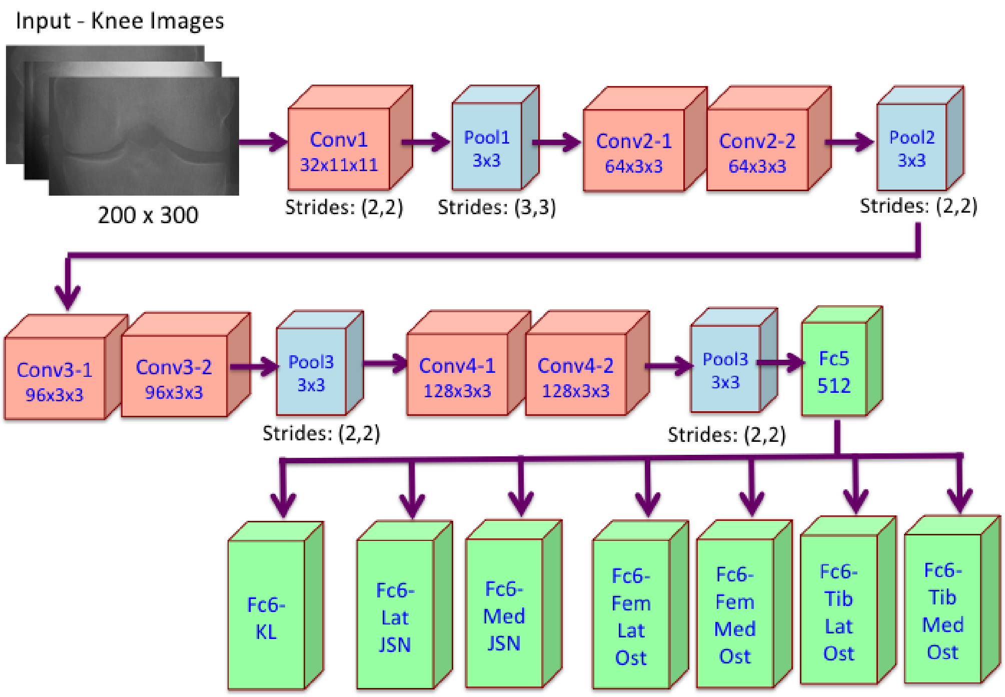


Figure 6.3: Initial configuration of the jointly trained network for classification of knee images based on KL grades, JSN, and osteophytes.

imbalance. Table 6.1 shows the grade-wise distribution of the knee radiographs as per KL grades, JSN, and osteophytes. The number of training and test samples is very low for JSN and osteophytes grades 2, 3 and 4 in comparison to grade 1. Table 6.9 shows the grade-wise precision, recall, and F_1 score of the jointly trained CNN for classifying femur medial osteophytes. These results are influenced by the number of training samples. For grade 0 the number of training samples are high (∼10,000) in total and the precision, recall, and F_1 score values are high (Table 6.9). Whereas for the other grades: grade 1 (∼3,000), grade 2 (∼1,000), and grade 3 (∼1,000), the number of training samples is low and the lack of sufficient training data is a reason for low balanced accuracy (42.1%) and also low precision, recall, and F_1 score for grade 1, 2, and 3 femur lateral osteophytes classification. These results are from the initial configuration. Next, the hyper-parameters of this network are tuned to investigate if there is an improvement in the classification

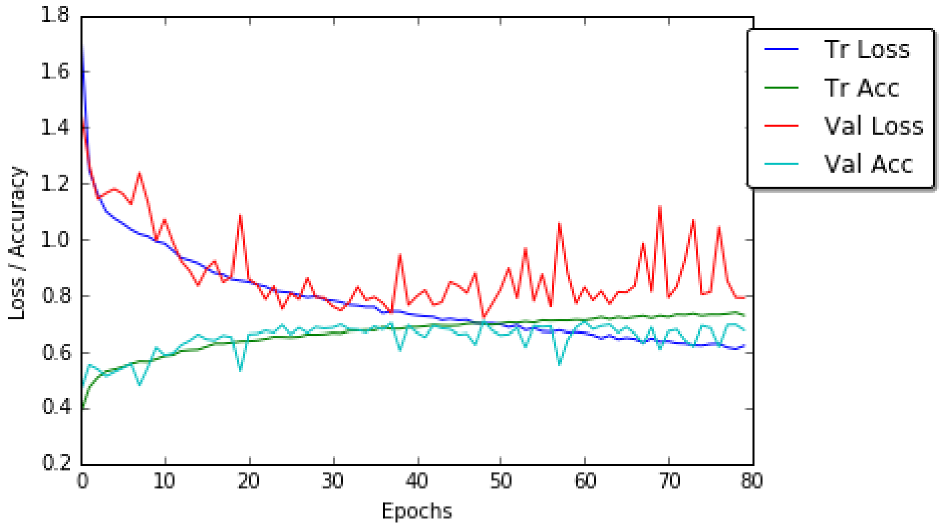


Figure 6.4: Learning curves for KL grades classification in jointly trained CNN.

Table 6.8: Classification results of jointly trained CNN to classify knee images based on KL grades, JSN, and osteophytes.

Variable	Acc.	Balanced Acc.	Precision	Recall	F_1 Score
KL grade	64.9%	63.4%	0.62	0.65	0.61
Lateral JSN	93.8%	70.2%	0.94	0.94	0.93
Medial JSN	79.0%	73.6%	0.78	0.79	0.78
Femur lateral osteophytes	72.7%	45.3%	0.67	0.73	0.68
Femur medial osteophytes	72.1%	46.3%	0.66	0.72	0.65
Tibia lateral osteophytes	74.4%	49.9%	0.72	0.74	0.72
Tibia medial osteophytes	63.1%	42.8%	0.61	0.63	0.6

performance.

6.5.3 Tuning the Hyper-parameters

The initial configuration (Table 6.7) contains eight layers of learned weights: seven convolutional layers and a fully connected layer. The hyper-parameters of this network are tuned by varying the number of convolutional layers, convolutional-pooling stages, number of convolutional kernels and kernel size. There is a slight decrease ($\sim$ 1 to 2%) in classification accuracy and other performance metrics on the validation set, when reducing the number of convolutional layers. Removing the conv4-1 and conv4-2 layers from the CNN

Table 6.9: Results of jointly trained CNN to classify femur medial osteophytes.

Grade	Precision	Recall	F_1 Score
0	0.78	0.97	0.87
1	0.37	0.12	0.19
2	0.31	0.08	0.12
3	0.50	0.51	0.50
Mean	0.66	0.73	0.67

(Table 6.7) decreases the classification accuracy of KL grades from 64.9% to 63.5%, and the balanced classification accuracy from 63.4% to 62%. There are no significant variations in the classification results of JSN and osteophytes.

Next, the number of convolutional layers are increased to check if there is any further improvement in the joint classification of KL grades, JSN, and osteophytes. A convolutional layer (conv5-1) is included after the maxPool4 layer to this CNN (Table 6.7) and this network is trained retaining the other previous settings. There is a small improvement in the overall classification results on increasing the depth of the network i.e. the number of layers with learned weights from eight to nine. The classification accuracy of KL grades increases from 64.9% to 65.5%, and the balanced accuracy increases from 63% to 63.6%. There is no further improvement in the classification results on increasing the number of convolutional layers. This also increases the number of free parameters in the network and leads to over fitting after 35 epochs, and also there is increase in total training time of the network for the same number of epochs.

6.5.4 Joint Training for Classification and Regression

In Section 5.4.3, it was shown that jointly training a CNN to optimise two loss functions: categorical cross-entropy for multi-class classification, and mean-squared error for regression of KL grades, giving better results in comparison to the individual networks trained for classification and regression. Motivated by this, a CNN is jointly trained for simultaneous classification and regression of JSN,

Table 6.10: Best performing jointly trained CNN for classification of knee images based on KL grades, JSN, and osteophytes.

Layer	Kernel	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2-1	64	3×3	1	64×49×74
conv2-2	64	3×3	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3-1	96	3×3	1	96×24×36
conv3-2	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4-1	128	3×3	1	128×11×17
conv4-2	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8
conv5-1	128	3×3	1	128×5×8
fc6	–	–	–	512
fc7-KL	–	–	–	5
fc7-LatJSN	–	–	–	4
fc7-MedJSN	–	–	–	4
fc7-FemLatOst	–	–	–	4
fc7-FemMedOst	–	–	–	4
fc7-TibLatOst	–	–	–	4
fc7-TibMedOst	–	–	–	4

osteophytes, and KL grades. A similar architecture is used from the previous approach (Table 6.7), except the last fully connected layer (fc6) includes seven multi-class classification outputs and seven regression outputs to quantify lateral and medial JSN, osteophytes in lateral and medial compartments of femur and tibia, and KL grades. The network is optimised with a weighted ratio of fourteen loss functions: seven classification losses (categorical cross-entropy) plus seven regression losses (mean-squared error). Different weight ratios are tested for the regression loss from 0.3 to 0.7. There is no improvement in the multi-class classification and regression results. Moreover, this joint training increases the complexity of the network by including more objectives to optimise the network.

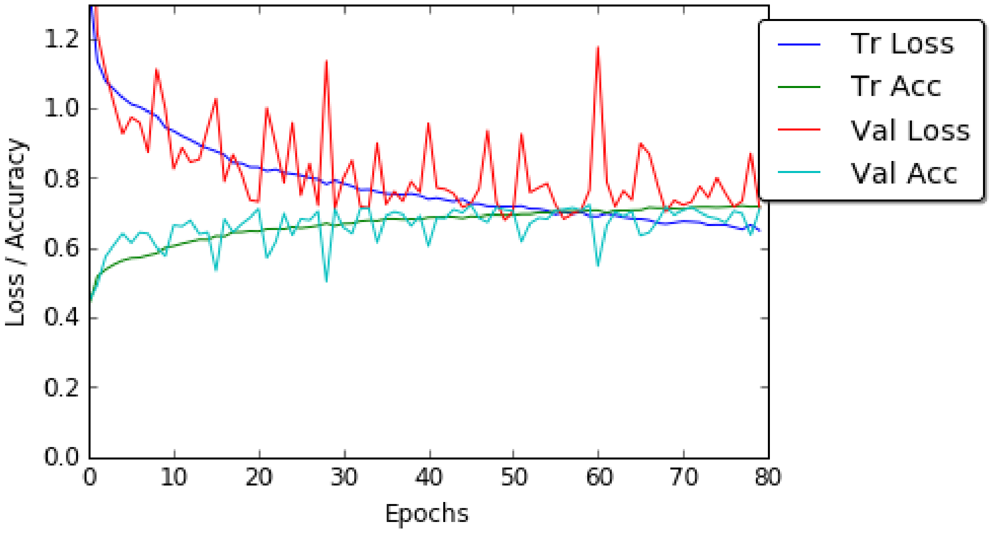
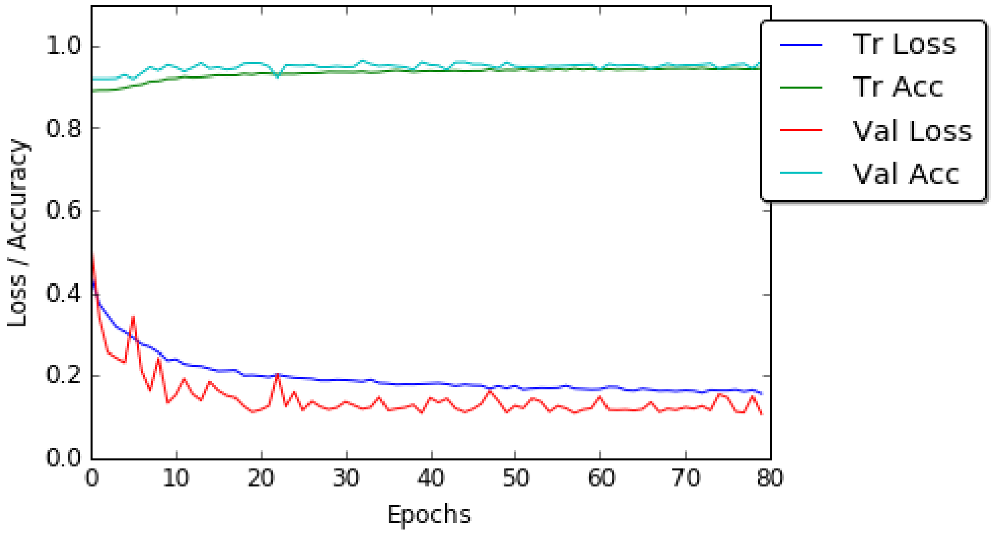


Figure 6.5: Learning curves for KL grades classification in jointly trained CNN.

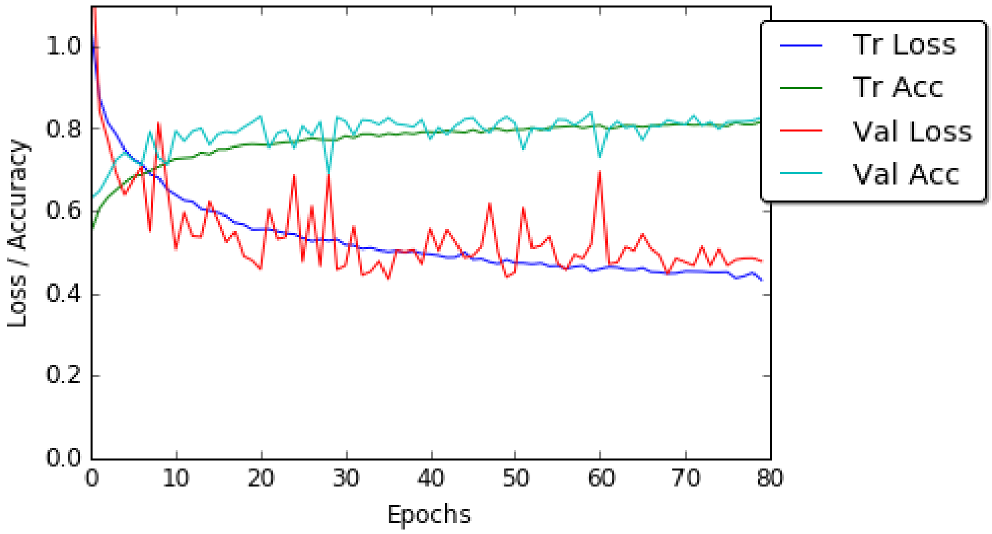
6.5.5 Best Performing Jointly trained CNN for Classification

An optimal configuration is reached after experimenting with different settings and the best performing jointly trained CNN (Table 6.10) is identified to classify the knee images based on KL grades, JSN, and osteophytes. This network contains ~3.1 million free parameters in total and nine layers of learned weights: eight convolutional layers and a fully connected layer. Each convolutional layer is followed by batch normalisation and ReLU activation. The last fully connected layer (fc7) contains seven multi-class classification outputs. To avoid over-fitting a drop out regularisation with a drop out ratio 0.3 is included after the fc6 layer, and L2 weight regularisation of 0.01 is included in all the convolutional and fully connected layers. This network is trained to optimise on the cumulative sum of seven categorical cross-entropy loss functions. The Adam optimiser is used with default parameters: initial learning rate (α) = 0.001, β_1 = 0.9, β_2 = 0.999, $\epsilon = 1\mathrm{e}^{-8}$ and train the network for 80 epochs.

The learning curves (Figure 6.5 and Figure 6.6) obtained whilst training the jointly trained network for KL grades and JSN classifications show convergence to minimum values of validation and training losses with consistent increase in the



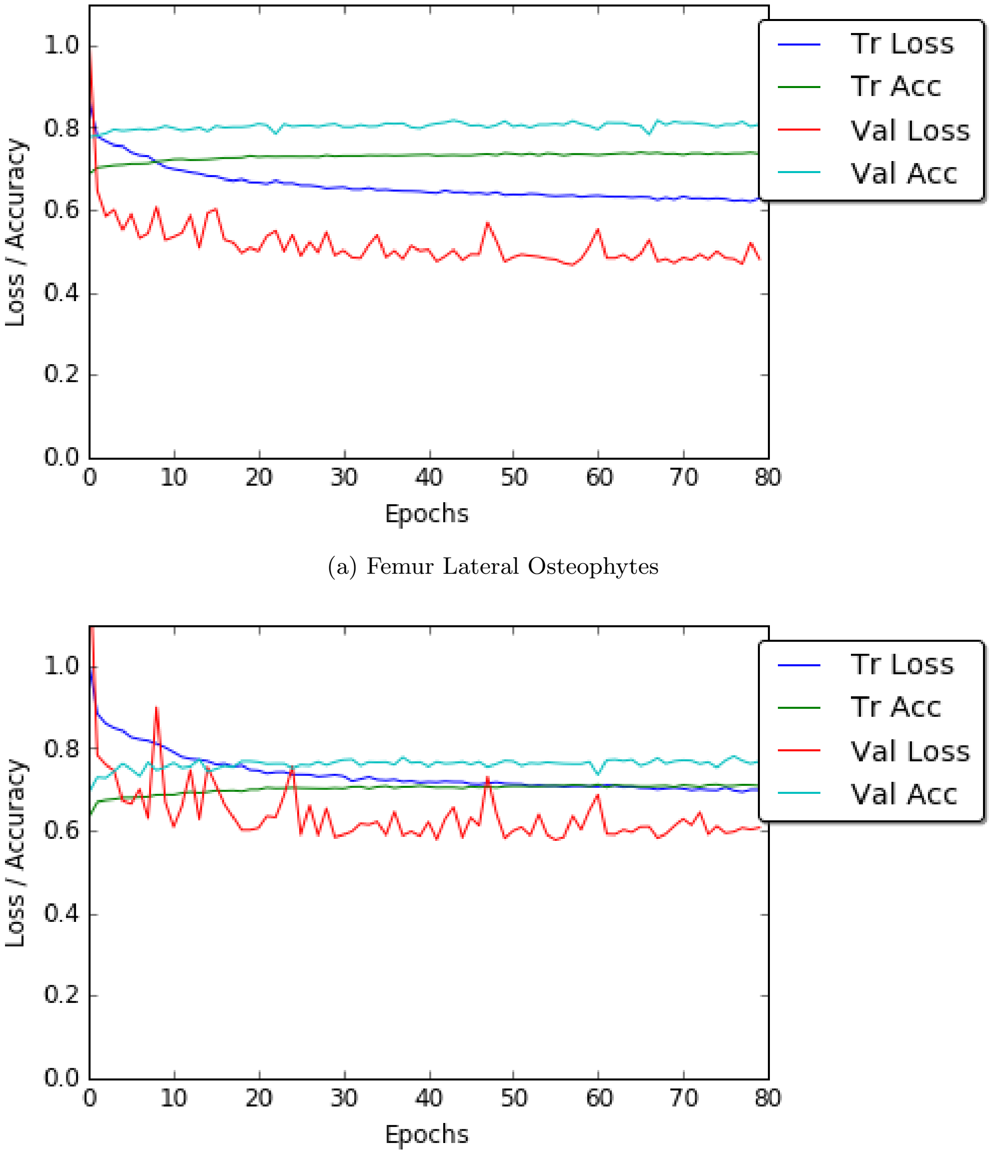
(a) Lateral JSN



(b) Medial JSN

Figure 6.6: Learning curves for classification of (a) lateral JSN and (b) medial JSN in jointly trained CNN.

training and validation accuracies. The learning curves for classifying osteophytes severity in femur (Figure 6.7) and tibia (Figure 6.8) show less variations in the training and validation losses, and the accuracies are reaching a plateau early in training. This shows the network is not able to learn an effective representation to discriminate osteophyte severity in the knee joints. The author suspects the main reason for this is the lack of sufficient training samples.



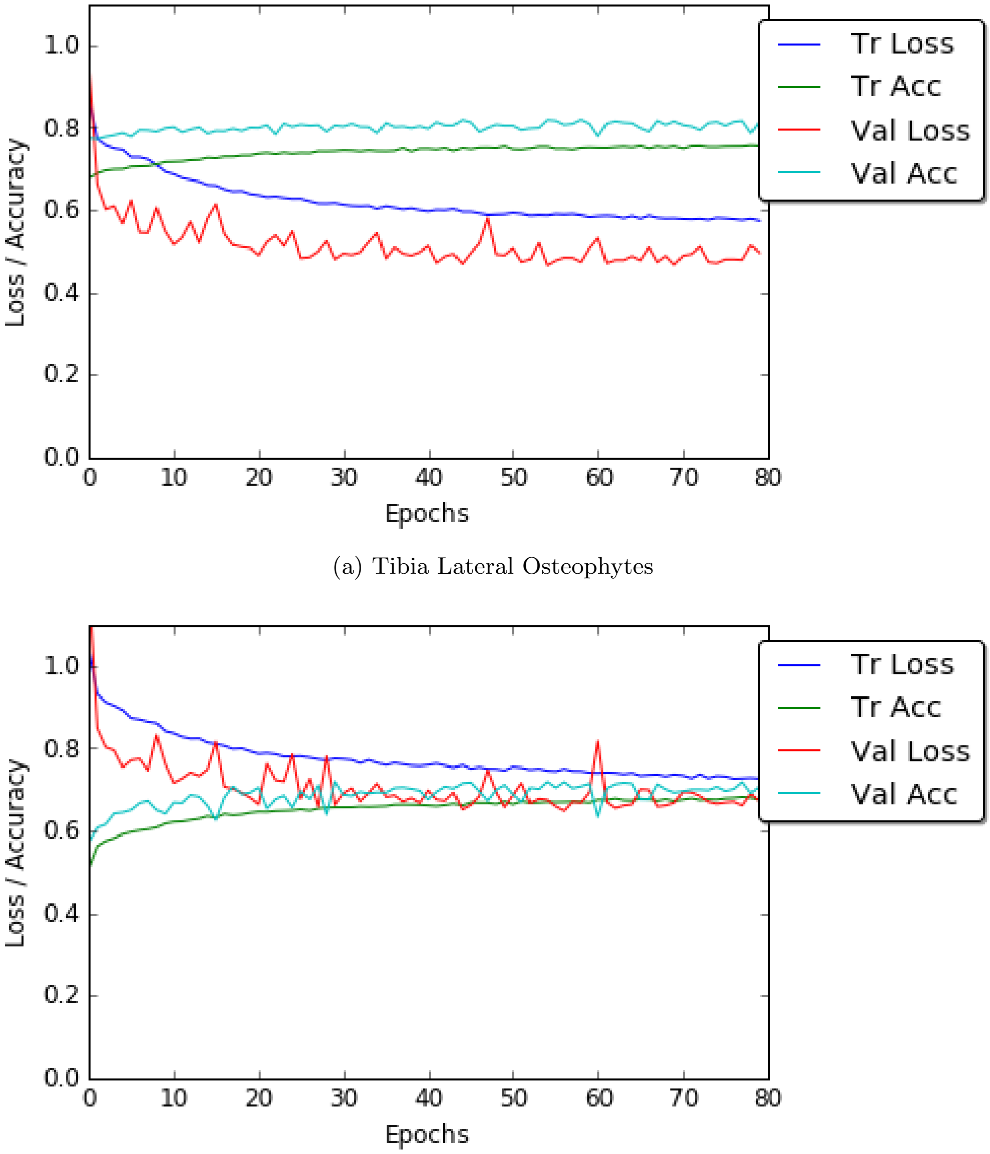
(a) Femur Lateral Osteophytes

(b) Femur Medial Osteophytes

Figure 6.7: Learning curves for classification of (a) femur lateral and (b) femur medial osteophytes severity in jointly trained CNN.

6.5.6 Jointly Trained CNN Classification Results

Table 6.11 shows the classification results: the multi-class classification accuracy, balanced accuracy, mean average precision, recall, and F_1 score of the best performing jointly trained CNN on the test data. The classification results show improvement in the multi-class classification accuracy and balanced accuracy for



(a) Tibia Lateral Osteophytes

(b) Tibia Medial Osteophytes

Figure 6.8: Learning curves for classification of (a) tibia lateral and (b) tibia medial osteophytes severity in jointly trained CNN.

KL grades, lateral and medial JSN in comparison to the previous results (Table 6.8). There is no significant improvement in the mean average precision, recall and F_1 Score for the KL grades and JSN in comparison to the previous results (Table 6.8). Even though the classification accuracy, precision, recall and F_1 Score are high for the osteophytes, the balanced accuracy is low. This is likely due to the dataset imbalance. Intuitively, the other reason for this could be the jointly

Table 6.11: Classification results of the best performing jointly trained CNN to classify knee images based on KL grades, JSN, and osteophytes.

Variable	Acc.	Balanced Acc.	Precision	Recall	F_1 Score
KL grade	65.5%	63.6%	0.63	0.65	0.60
Lateral JSN	93.9%	69.1%	0.93	0.94	0.93
Medial JSN	78.1%	73.4%	0.77	0.78	0.75
Femur lateral osteophytes	73.4%	44.3%	0.65	0.73	0.67
Femur medial osteophytes	71.8%	47.9%	0.67	0.72	0.66
Tibia lateral osteophytes	74.8%	47.6%	0.71	0.75	0.72
Tibia medial osteophytes	64.0%	45.8%	0.62	0.64	0.61

trained network is not able to discriminate the small anatomical variations due to osteophytes formation in the knee joints.

6.6 Comparison of Individual and Joint Training Results

The results of the jointly trained CNN are compared to the results of the individually trained CNNs to classify KL grades, JSN, and osteophytes. The results show (Table 6.12) high multi-class classification accuracies for the jointly trained CNN in comparison to the individual CNNs. There is an improvement in the balanced classification accuracy of the KL grades from 59% (individual training) to 63.6% (joint training). Plus there is an improvement in the balanced classification accuracy of the lateral JSN from 48.6% to 69.1%, and medial JSN from 72.3% to 73.4%. These results shows that jointly training a CNN following multi-objective convolutional learning is highly effective in comparison to individually training CNNs to classify knee images based on KL grades and JSN. However, there is no significant improvement in the balanced classification accuracies of osteophytes in the joint training. This is mainly due to the lack of sufficient training samples for severe grades of osteophytes and the imbalance in the datasets. Due to the lack of training data, the network is not learning an

Table 6.12: Comparing the results from individually trained CNNs to the jointly trained CNN for classifying knee images based on KL grades, JSN, and osteophytes.

Variable	Individual CNNs		Jointly Trained CNN	
	Acc.	Balanced Acc.	Acc.	Balanced Acc.
KL grade	61.8%	59.0%	**65.5%**	**63.6%**
Lateral JSN	90.8%	48.6%	**93.9%**	**69.1%**
Medial JSN	76.5%	72.3%	**78.1%**	**73.4%**
Femur lateral osteophytes	70.3%	**44.7%**	**73.4%**	44.3%
Femur medial osteophytes	70.8%	47.4%	**71.8%**	**47.9%**
Tibia lateral osteophytes	70.2%	45.3%	**74.8%**	**47.6%**
Tibia medial osteophytes	57.3%	**46.0%**	**64.0%**	45.8%

effective representation to discriminate the severity of osteophytes formation in the knee. In the next section, more evidence is provided to support this argument by using binary classification of knee images to detect knee OA.

6.7 Comparison of Classification Results to the OAI Reliability Readings

The automatic knee OA assessment results are compared to the radiologic outcomes from the OAI: specifically project 15 test-retest reliability of semi-quantitative readings from knee radiographs. In project 15, the OAI reading centre used 150 participants (300 knee joints) to evaluate the reliability of the KL scores and the OARSI central reading scores. There is a declaration quoting that this sample was representative of the entire cohort with respect to the grades of knee OA, radiographic progression and incidence, and the number of time points with radiographs for a participant. Two separate readings were used to retest the original readings and the readers were blinded to the original scores of the readings. Simple kappa coefficientswere used to evaluate agreement between the two readings when the variable was dichotomous. Weighted kappa coefficients were used if the variable had more than two ordinal categories.

The OAI provide kappa coefficients with 95% confidence interval (CI) for the agreement between the first and the second readings of the KL grades and all the distinct knee OA features. The reliability readings from the OAI can be classified into two: 1) Weighted kappa values for multi-class classification of KL grades in an ordinal scale of (0–4), and for JSN and osteophytes in a scale (0–3), 2) Simple Kappa values for binary classification of KL grades in a binary scale such as KL grade < 2 is 0 and KL grade ≥ 2 is 1, and for any JSN, or any osteophytes < 1 is 0; for grades (1–3) is 1.

6.7.1 Multi-class Classification

The weighted kappa values are calculated for the multi-class classification results from the jointly trained CNN (Table 6.10) and compare this to the gold standard: the weighted kappa values from the OAI reliability readings based on an ordinal scale. Figure 6.9 and Table 6.13 shows the kappa values with 95% CI on the jointly trained CNN for multi-class classification and the radiologic reliability readings from the OAI, and the kappa values with 95% CI for the assessments based on an ordinal scale. There is an overlap in the error bars (CI) of the classification results and the OAI reliability readings. From the results it is evident that our results (predictions) agree with the gold standard as well as the annotators or evaluators who produced the gold standard agree with one another.

Unfortunately, the results for osteophytes classification show low kappa values in comparison to the OAI reliability readings. The reason for this is lack of sufficient training data belonging to higher grades (1–3) of osteophytes. Due to scarce training data, the network is not able to learn an effective representation to discriminate the small anatomical variations for osteophytes formation in the knee joints. The classification of osteophytes can perhaps be improved by increasing the number of training samples and investigating other CNN configurations.

Table 6.13: Comparison of the multi-class classification results to the OAI radiologic reliability readings in ordinal scale.

Variable	Multi-class Clsf. Kappa (95% CI)	OAI Kappa (95% CI)
KL grade	0.69 (0.68 – 0.71)	0.70 (0.65 – 0.76)
Lateral JSN	0.80 (0.77 – 0.83)	0.87 (0.76 – 0.98)
Medial JSN	0.75 (0.73 – 0.77)	0.75 (0.68 – 0.81)
Femur lateral ost.	0.47 (0.44 – 0.50)	0.69 (0.60 – 0.79)
Femur medial ost.	0.61 (0.59 – 0.63)	0.73 (0.66 – 0.81)
Tibia lateral ost.	0.52 (0.49 – 0.54)	0.70 (0.61 – 0.79)
Tibia medial ost.	0.48 (0.46 – 0.51)	0.69 (0.60 – 0.77)

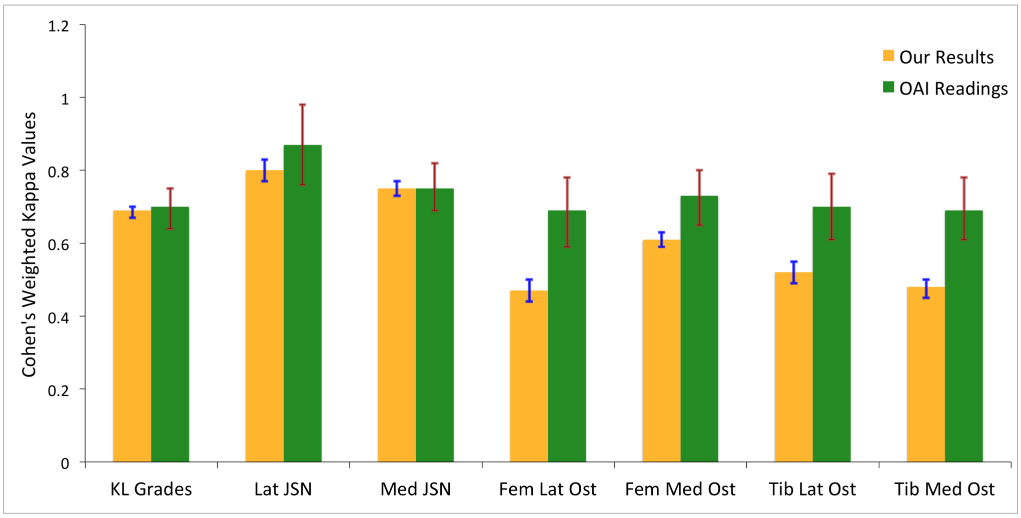


Figure 6.9: Comparison of the results to the OAI radiologic reliability readings.

6.7.2 Binary Classification

The same network configuration (Table 6.10) that gave the previous best results for multi-class classification is used to jointly train a CNN for binary classification of knee images based on KL grades, JSN, and osteophytes. The same strategy followed by the OAI is used to prepare the labels to train the network for binary classification of KL grades (KL grade < 2 is 0 and KL grade ≥ 2 is 1), and for binary classification of JSN and osteophytes (any variable < 1 is 0; otherwise 1). The fc6 layers are replaced with 2 outputs for binary classification and change the loss functions to binary cross-entropy. The network is trained for 80 epochs using

Table 6.14: Comparison of the binary classification results to the OAI radiologic reliability readings in binary scale.

Variable	Binary Clsf. Kappa (95% CI)	OAI Kappa (95% CI)
KL grade	0.68 (0.65 – 0.70)	0.70 (0.62 – 0.78)
Lateral JSN	0.80 (0.76 – 0.83)	0.83 (0.70 – 0.96)
Medial JSN	0.67 (0.65 – 0.70)	0.74 (0.66 – 0.82)
Femur lateral ost.	0.50 (0.47 – 0.54)	0.78 (0.70 – 0.85)
Femur medial ost.	0.60 (0.58 – 0.64)	0.78 (0.70 – 0.85)
Tibia lateral ost.	0.57 (0.54 – 0.60)	0.71 (0.63 – 0.79)
Tibia medial ost.	0.52 (0.49 – 0.55)	0.71 (0.63 – 0.79)

the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\mathrm{e}^{-8}$.

Table 6.14 shows the the kappa values with 95% CI for the binary classification and the radiologic reliability readings from the OAI for the assessments based on a binary scale. Like the multi-class classification, the binary classification results are also on par with the reliability readings and agrees with the gold standard. However, for the osteophytes classification the kappa values are low in comparison to the gold standard. This is again due to the lack of sufficient training data.

6.7.3 Comparison of Binary and Multi-class Classifications

The results of binary and multi-class classifications (Table 6.15) from the same network configuration are compared to show that more training data can improve the classification of KL grades, JSN, and osteophytes. From the results it is evident that both the classification accuracy and balanced accuracy is high for binary classification in comparison to multi-class classification. Even the kappa values for binary classification (Table 6.14) of osteophytes is high in comparison to the multi-class classification. This emphasises the fact that sufficient training data is essential to learn an effective representation and the predictions of the network to generalise well. The assessments on a binary scale are useful to detect if there is any joint space narrowing between the femur and tibia, presence of osteophytes in

Table 6.15: Comparing the results of multi-class classification to binary classification for classifying knee images based on KL grades, JSN, and osteophytes.

Variable	Multi-class Classification		Binary Classification	
	Acc.	Balanced Acc.	Acc.	Balanced Acc.
KL grade	65.5%	63.6%	83.6%	84.1%
Lateral JSN	93.9%	69.1%	96.2%	87.2%
Medial JSN	78.1%	73.4%	84.3%	82.8%
Femur lateral ost.	73.4%	44.3%	80.7%	72.8%
Femur medial ost.	71.8%	47.9%	82.4%	78.6%
Tibia lateral ost.	74.8%	47.6%	82.4%	76.9%
Tibia medial ost.	64.0%	45.8%	75.9%	76.3%

femur or tibia and to detect knee OA incidence.

In summary, the jointly trained CNN achieves high binary classification accuracy for detecting the incidence of knee OA, and presence of JSN and osteophytes. The multi-class classification results are on par with the inter-rater agreement to quantify knee OA severity based on KL grades and JSN.

6.8 An Automatic Knee OA Diagnostic System

In this chapter, two approaches are presented to quantify knee OA severity: 1) training CNNs separately to classify knee images based on KL grades, JSN, and osteophytes, and 2) jointly training a CNN using multi-objective convolutional learning for simultaneous classification of KL grades, JSN, and osteophytes. The classification results (Table 6.12) show that the jointly trained network outperforms the individual CNNs to classify KL grades, JSN, and osteophytes. An automatic knee OA diagnostic system is developed combining the automatic localisation pipeline (Section 4.3) and the quantification pipeline developed in the previous section.

Figure 6.10 shows the proposed end-to-end diagnostic pipeline to automatically quantify knee OA severity based on KL grades and the knee OA diagnostic features: JSN and osteophytes. The input X-ray images are subjected to

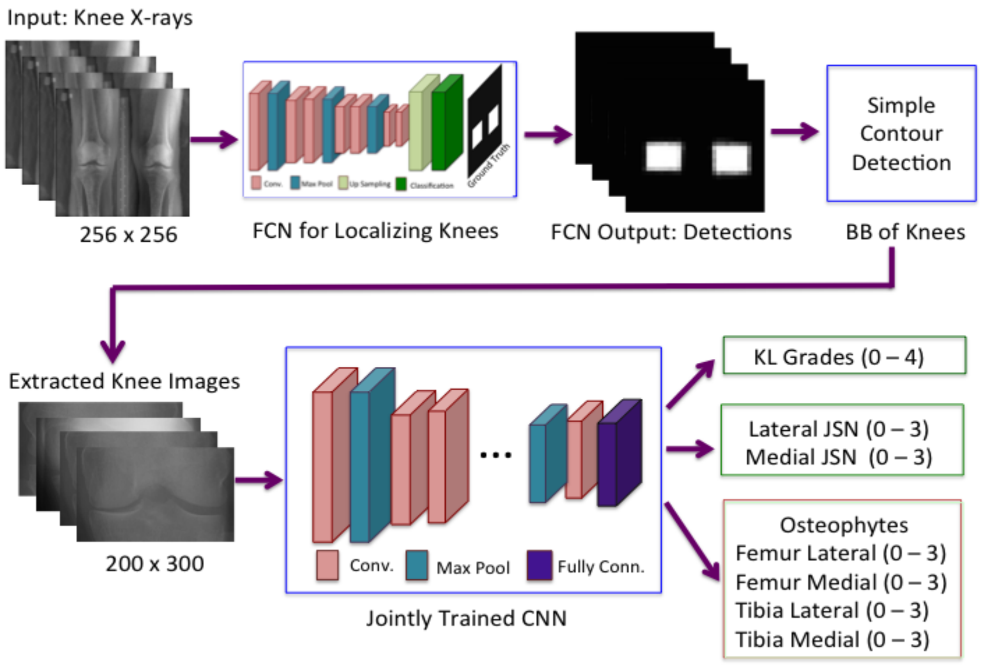


Figure 6.10: An end-to-end knee OA diagnostic pipeline.

histogram equalisation, mean normalisation and resized to a fixed size 256×256. A fully convolutional network (FCN) is used to automatically detect the ROI, the knee joint regions. The bounding box coordinates of the ROI are calculated using simple contour detection. The knee joint regions are extracted from the knee radiographs using the bounding box coordinates. The localised and extracted knee images are resized to 200×300 to preserve the mean aspect ratio (∼1.6) and fed to the jointly trained CNN (Table 6.10). This CNN gives seven multi-class classification outputs in total based on KL grades, lateral and medial JSN, osteophytes severity of femur and tibia in the lateral and the medial compartments.

The major pathological features that indicate the onset of knee OA include: reduction in joint space width due to loss of knee cartilage, and the formation of bone spurs (osteophytes) or bony projections along the joint margins. The author believes that quantifying these features along with the KL grades can provide deeper insights to assess knee OA severity and to study the progression of knee

OA. Therefore, a deep learning-based automatic knee OA diagnostic system that can provide simultaneous predictions of KL grades, JSN, and osteophytes is developed.

6.9 Chapter Summary

Two approaches are presented in this chapter to automatically quantify the distinct knee OA diagnostic features, JSN and osteophytes based on the OARSI central reading scores. First, CNNs are trained separately to classify the knee images based on the lateral and medial JSN, femoral and tibial osteophytes in lateral and medial compartments. Next, CNNs are jointly trained following the multi-objective convolutional learning approach to quantify JSN and osteophytes along with the KL grades. The rationale behind this is providing multiple ground truths to the network can improve the overall quantification results. The jointly trained CNN gave better classification results in comparison to the individually trained CNNs. High classification accuracy for KL grades and JSN classification was achieved. However, classifying osteophytes is challenging due to scarce data in the higher grades of osteophytes.

The automatic quantification results are compared to the gold standard: the reliability readings from the OAI. Kappa coefficients were used to evaluate inter-rater agreement. The classification results are on par with the inter-rater agreement and reliability readings (kappa values) to assess the KL grades, JSN, and osteophytes in an ordinal scale. Classification results for osteophytes, however, are below par to the kappa values. High binary classification accuracies to assess the presence of JSN and osteophytes, and to detect the incidence of knee OA are achieved.

To conclude this chapter, an automatic knee OA diagnostic system is presented combining the localisation] pipeline based on the FCN (Section 4.3) and the classifier based on the jointly trained CNN (Section 6.5) to quantify KL grades, JSN, and osteophytes.

Chapter 7

Conclusion

7.1 Thesis Overview

The main goal of this thesis is to advance the state-of-the-art in computer aided diagnostics of the severity of knee OA by developing deep learning based automatic methods. According to the literature, automatic assessment of knee OA severity has been previously approached as an image classification problem and existing approaches report low accuracy for multi-class and classification of successive grades. The state-of-the-art machine learning based methods are investigated for image classification, and developed new methods using convolutional neural networks (CNNs) to automatically classify knee OA images. A significant outcome of this thesis is a new automatic knee OA diagnostic system that achieves high accuracy, on par with radiologic reliability readings, which are considered the gold standard for knee OA assessment.

A summary of the investigations, research findings, experimental results, and the proposed solutions in this thesis is as follows.

Chapter 1 introduced knee OA degradation and knee OA diagnostic features, discussed the clinical significance of knee OA research, described the motivations for this thesis, presented the hypotheses, the research questions and the research objectives, and outlined the structure of this thesis.

Chapter 2 reviewed the literature in knee OA assessment, computer aided diagnostics of knee OA and the diagnostic features, and introduced necessary technical background. According to the literature, an automatic knee OA diagnostic pipeline consists of two steps: localising the knee joints and quantifying the OA severity in the localised knee joints. The previous approaches for detecting knee joints in X-ray images can be classified into manual, interactive, and automatic methods. The drawbacks of manual and interactive methods, and the advantages of automatic methods are outlined. It was concluded that according to the literature the automatic localisation of knee joints in radiographs still remains a challenge.

The previous approaches in the literature for assessing radiographic knee OA severity were classified into interactive and automatic methods. The interactive methods are objective and accurate, but a great deal of manual intervention is required and these methods may become laborious and time-consuming for a large number of subjects. Some of these drawbacks are overcome in automatic methods. However, existing approaches achieve low multi-class classification accuracy and classifying successive grades remains a challenge.

Chapter 3 presented the baseline approaches and the preliminary experiments to automatically localise knee joints in radiographs and to classify the localised knee joints. Template matching was implemented as a baseline for the automatic localisation of knee joints. The computations in template matching are mainly based on intensity-level differences and this method gave low precision ($\sim 30\%$) for detecting the knee joints. A new method based on a SVM for automatic detection of knee joints was proposed. This method improved the results with a detection precision above 80%. The reason for improved results in this method is due to the use of Sobel horizontal image gradients and horizontal edge based discrimination, unlike simple intensity-level discrimination used in the template matching method.

In Chapter 3, two approaches were explored for classifying knee images: 1) using hand-crafted features with conventional classifiers such as SVM, kNN and SVR, and 2) learning feature representations in a supervised manner using a CNN. First, the state-of-the-art hand-crafted features were tested to classify knee images. From the experimental results it was concluded that the classification methods using hand-crafted features and conventional classifiers can be further improved by including more effective features. Next, supervised feature learning using off-the-shelf CNNs such as the VGG-16 network was investigated to classify knee images and promising results were achieved. The outcomes of the baseline methods in this chapter motivated the use of deep learning methods for automatic localisation of knee joints and classification of the localised knee joints to quantify OA severity.

Chapter 4 focused on developing automatic methods to localise knee joints in radiographs using supervised feature learning. Two new approaches were introduced for automatically detecting and localising the knee joints in X-ray images using full convolutional networks (FCNs). First, FCNs were trained to automatically detect the knee joint centres. A fixed size region was cropped as the ROI with reference to the detected centres. Even though this approach achieved high detection accuracy, the extracted knee joints had some limitations due to the variations in the resolution of the knee radiographs and the anatomical variations of the knees. Next, FCNs were trained to directly localise the ROI instead of knee joint centres. The objective was to overcome the drawbacks in the previous approach and to further improve the localisation accuracy. Near perfect localisation results were achieved and the experimental results outperformed the previous methods to localise the knee joints.

Chapter 5 focused on automatically quantifying knee OA severity using CNNs. Four approaches were presented for the classification, regression and ordinal regression of knee OA images based on KL grades. First, the existing pre-trained CNNs were investigated for fixed feature extraction and classified the knee OA images using CNN features and linear SVMs. Next, the pre-trained CNNs were fine-tuned using the transfer learning approach. The predictions or outputs from these methods are ordinal KL grades (0, 1, 2, 3, or 4). The author argued that quantifying knee OA severity in a continuous scale (0–4) is more appropriate as the OA degradation is progressive in nature, not discrete. Regression was used to quantify the knee OA severity in a continuous scale. The classification and regression results from the methods proposed in this chapter outperformed the previous best results.

In the second approach, CNNs were trained from scratch for classification and regression of knee images. The objective was to further improve the quantification results. A Lightweight architecture was used with fewer (∼4 to ∼5 million) free parameters in the CNNs as the training data is relatively scarce. The fully trained CNN for classification achieved high classification accuracy in comparison to the pre-trained CNNs. The fully trained CNN for regression did not achieve good results as no ground truth of KL grades in a continuous scale is available.

In the third approach, CNNs were fully trained using multi-objective learning for simultaneous classification and regression. The author argued that optimising a CNN with two loss functions provides a stronger error signal and was a step towards improving the overall quantification, considering both classification and regression results. The jointly trained CNN achieved better quantification results with a high classification accuracy in comparison to the previous methods.

As the last approach, CNNs were fully trained for ordinal regression. This method achieved low mean-squared error and outperformed the other methods for assessing knee OA severity in a continuous scale. In summary, A progressive improvement was achieved in the quantification performance with increase in classification accuracy and other performance metrics. This chapter was concluded with an error analysis that discussed the possible reasons for the misclassification from the jointly trained CNN.

Chapter 6 presented two new approaches to automatically quantify distinct knee OA features: specifically JSN and osteophytes. First, CNNs were trained separately to classify the knee images based on the lateral and medial JSN, femoral and tibial osteophytes in lateral and medial compartments. Next, CNNs were jointly trained following the multi-objective convolutional learning approach to quantify JSN and osteophytes along with the KL grades. The author believed that providing multiple ground truths to the network could improve the overall quantification results. The jointly trained CNN gave better classification results in comparison to the separately trained CNNs. High classification accuracy for KL grades and JSN classification was achieved. Nevertheless, classifying osteophytes was challenging due to scarcity of training data in the higher grades of osteophytes. The automatic quantification results were compared to the reliability readings from the OAI, which is considered the gold standard for knee OA assessment. The classification results are on par with the inter-rater agreement and reliability readings (kappa values) to assess the KL grades, and JSN. Unfortunately, the classification results for osteophytes are poor in comparison to the gold standard. This chapter concluded presenting an automatic knee OA diagnostic system combining the localisation pipeline based on the FCN (developed in Chapter 4) and the classifier based on the jointly trained CNN (Section 6.5) to quantify KL grades, JSN, and osteophytes.

7.2 Research Questions and Proposed Solutions

This thesis investigated automated methods to assess knee OA severity and the diagnostic features based on six research questions in conjunction with four hypotheses. The hypothesis and research questions are revisited, and these are examined with respect to the proposed solutions and the experimental results.

H1. *Learning feature representation and classification using supervised deep learning is more effective for assessing the severity of knee OA than conventional classification using hand-crafted features.*

H2. *Evaluating the automatic knee OA predictions using a continuous distance-based metric like mean squared error instead of classification accuracy is more appropriate and KL grades predictions can be approached as a regression problem. Consequently, training a CNN for optimising a weighted ratio of two loss functions for simultaneous classification and regression can improve the accuracy of quantifying* *knee OA severity.*

H3. *Jointly training a CNN for quantifying the clinical diagnostic features of knee OA such as joint space narrowing (JSN) and osteophytes, along with the KL grades will improve the overall quantification of knee OA severity.*

H4. *As a result, a highly accurate computer aided diagnostic system can be built to assess knee OA.*

From the hypotheses, the following research questions are outlined.

H1 - RQ1. What is the most efficient method for localising the ROI: the knee joint regions in X-ray images, in terms of speed and accuracy that also supports feature learning and classification using CNNs?

A new FCN based pipeline was developed to automatically localise the knee joint regions in X-ray images. This research question is linked to hypothesis H1, and is investigated in Chapters 3 and 4. The investigation included the following methods. As a baseline approach, template matching was implemented for detecting knee joints in Section 3.3.1. and this method was not precise. A new SVM-based method using Sobel horizontal image gradients was proposed in Section 3.3.2, in an attempt to improve the precision of knee joint detection. This method gave better results, but falls short of perfect detection. Two new approaches were investigated for localising the knee joint regions in Chapter 4. FCNs were trained to directly localise the knee joint region in Section 4.3 and this method achieved near perfect localisation results. Later, this localisation pipeline was combined with the quantification pipeline (Chapter 6) to develop an automatic knee OA diagnostic system.

H1 - RQ2. Instead of using hand-crafted features, is it possible to learn effective feature representations using a supervised deep learning method, in

particular a convolutional neural network (CNN), for efficient and accurate fine-grained classification of knee OA images?

This research question is also linked to hypothesis H1, and is investigated in Chapter 3 and 5. Supervised feature learning using CNNs for efficient classification of knee OA images was investigated in Section 3.4.3, instead of using hand-crafted features. The off-the-shelf pre-trained CNNs were fine-tuned through transfer learning in Section 5.2. The author argued that the knowledge transfer may be limited by the substantial differences between the source and the target domains, and can mitigate the performance of the fine tuned CNNs. Therefore, as the next approach CNNs were trained from scratch specific to this task instead of fine-tuning the existing CNNs in Section 5.3. Better classification results in comparison to the previous methods were achieved.

H2 - RQ3. As knee OA is progressive, can the categorisation of knee OA images be approached as a regression problem instead of classification?

This research question is linked to hypothesis H2, and is investigated in Chapter 5. The pre-trained CNNs were fine-tuned using both classification loss and regression loss in Section 5.2. It was shown that the CNN fine-tuned with regression loss improves the classification accuracy in comparison to the CNN fine-tuned with classification loss in Section 5.2.5. CNNs were trained from scratch in order to improve the regression results in Section 5.3.8. Better results were achieved for regression of knee OA images in the jointly trained network for classification and regression in Section 5.4.4.

H2 - RQ4. an a CNN trained with a weighted ratio of two loss functions such as categorical cross entropy and mean squared error improve the assessment of knee OA severity?

This research question is linked to hypothesis H2, and is investigated in Chapter 5. A CNN was jointly trained for simultaneous classification and regression of knee OA images in Section 5.4. It was shown that there was an improvement in the classification performance in this jointly trained CNN in comparison to the CNN only trained for classification in Section 5.4.5.

H2 - RQ5. Can ordinal regression be applied to automatically assess knee OA severity? How does this improve the overall assessment of knee OA severity?

This research question is linked to hypothesis H2, and is investigated in Chapter 5. The jointly trained CNN configuration was modified to perform ordinal regression. A CNN was trained with a softmax dense layer as a hidden layer, applying linear weights to the classification output followed by a dense layer with linear activations for regression (Section 5.5). It was shown that the ordinal regression improves the assessment of knee OA severity in a continuous scale in Section 5.5.3.

H3 - RQ6. Can joint training a CNN for quantifying knee OA clinical features such as JSN and osteophytes along with KL grades further improve the overall quantification of knee OA severity?

This research question is linked to hypothesis H3, and is investigated in Chapter 6. First, CNNs were trained from scratch to quantify lateral and medial JSN individually in Section 6.3, to quantify femoral and tibial osteophytes in lateral and medial compartments in Section 6.4. Jointly trained CNNs were investigated for quantifying JSN and osteophytes along with the KL grades to explore if there is any further improvement in the overall quantification of knee OA in Section 6.5. It was concluded that the joint training achieves better quantification results in comparison to the individual trainings in Section 6.6.

H4 - RQ7. How well do the results agree with the gold standard for assessing knee OA? Can the proposed methods be applied in practical computer aided diagnosis (CAD) of knee OA?

This research question is linked to hypothesis H4, and is investigated in Chapter 6. An automatic knee OA diagnostic system was developed combining the FCN for localising the knee joints and the CNN jointly trained for quantifying knee OA severity in Section 6.8. The performance of knee OA diagnostic system was compared to the radiologic reliability readings from the OAI in Section 6.7. Cohen's weighted kappa values were used for the comparison of the results. The quantification results for KL grades and JSN are on par with the OAI reliability readings, which is considered the gold standard for knee OA assessment. In conclusion, the proposed methods in this thesis can be used as a supporting system to assess knee OA severity and to study the progression of knee OA.

7.3 Research Contributions

The research contributions of this thesis are as follows.

1. Proposing a novel and highly accurate technique to automatically detect and localise the knee joints from the X-ray images using a fully convolutional network (FCN).

2. Developing a classifier based on a CNN to assess knee OA severity that is highly accurate in comparison to existing methods.

3. Proposing a novel approach to train a CNN with a weighted ratio of two loss functions: categorical cross entropy and mean squared error with the natural benefit of predicting knee OA severity in ordinal (0,1,2,3, and 4) and continuous (0–4) scales.

4. Developing an ordinal regression approach using CNNs to automatically quantify knee OA severity in a continuous scale.

5. Developing CNN classifiers to assess the distinct knee OA diagnostic features: lateral and medial JSN, femoral and tibial osteophytes in lateral and medial

compartments.

6. Further improving the quantification of knee OA severity by jointly training CNN to predict the knee OA clinical features along with the KL grades.

7. Developing an automatic knee OA diagnostic system i.e. an end-to-end pipeline incorporating the FCN for automatically localising the knee joints and the CNN for automatically quantifying the localised knee joints. The overall assessment of knee OA by this system agrees with the gold standard.

7.4 Future Work

There are several potential directions for future work and further development of the research in this thesis. Some of the interesting extensions and prospects are outlined in the following.

Training an end-to-end deep learning model: The knee OA diagnostic pipeline consists of two steps: 1) localising the knee joints in radiographs and 2) assessing the knee OA severity from the localised knee joints. In this thesis, a FCN was trained for automatic localisation and a CNN was jointly trained for classification and regression of knee OA images. It would be interesting to train a single deep learning model integrating the FCN for localisation and the CNN for classification and/or regression, as this would further improve the automatic assessment of knee OA. Recently, end-to-end trained CNNs have become highly successful in saliency prediction [113], object detection [114], video classification [115], text recognition [60], and speech recognition [116].

Using semantic segmentations to measure joint space width: Among the knee OA diagnostic features, JSN is highly sensitive to changes due to disease progression. The proposed approach to automatically localise the knee joints using fully convolutional network can be extended for semantic segmentation of the knee joints and can be used to automatically measure the joint space width (JSW) between the femur and tibia. Pixel level knee joint annotations in radiographs are needed to measure the JSW.

Assessing the progression of knee OA severity: The automatic quantification methods developed in this thesis can be extended to assess the progression and early detection of knee OA severity. The baseline datasets are used from the OAI and the MOST dataset. Datasets are available for annual follow-up visits up to 9 years. These datasets could be used to detect the features predictive of radiographic knee OA progression. Shamir et al. reported a similar approach using WNDCHRM that predicted whether a knee would change from KL grade 0 to grade 3 with 72% accuracy using 20 years of data [2].

Relating the automatic quantification results to knee pain: The primary clinical features to assess knee OA are radiographic evidence of deformity and pain [117]. It would be interesting to study the relationship between the automatic assessments of the proposed methods (KL grades) to WOMAC scores for knee pain. WOMAC is one among the most widely used assessments in knee OA.

Relating the automatic quantification results to physiological variables: There are several pathological and physiological variables available in the OAI and the MOST datasets. These variables include potential predictors of knee pain status. It would be interesting to study the relationship between the outcomes of the automatic methods and the predictions from the pathological and physiological variables.

Investigating human level accuracy: The radiologic reliability readings from the OAI used 150 participants (300 knees) to evaluate the test-retest reliability of semi-quantitative readings. This is considered the current gold standard for knee OA assessment. Simple kappa coefficients and weighted kappa coefficients were used to evaluate the inter-rater agreement. Investigating the human level accuracy for a large sample and comparing it with the automatic quantification results would provide insight to help reduce the error involved in automatic assessments.

7.5 Concluding Remarks

In recent years, deep learning-based approaches, in particular convolutional neural networks have become highly successful in many computer vision tasks and medical applications. This thesis mainly focused on developing a deep learning-based computer aided diagnostic system. The proposed approaches in this thesis are related to two main medical applications: localising or automatically detecting and extracting a region of interest (ROI) from a radiograph, and classifying the ROI to automatically assess disease severity. The FCN-based localisation approach could be extended to other medical applications such as localising a substructure or a ROI in MRI and CT scan images, object or lesion detection, and locating anatomical landmarks or identifying imaging markers to study the disease progression. For instance, a similar FCN-based approach was followed to automatically detect and quantify ischemic injury (brain lesions) on diffusion-weighted MRI of infants, and the state-of-the-art was improved by achieving promising results.

In the author's opinion the most interesting research findings in this thesis are as follows. First, fine-tuning off-the-shelf CNNs pre-trained on very large datasets such as ImageNet (with ∼1M images) to classify knee images with relatively small datasets (with ∼10,000 images) is promising for medical image classification. The main challenge in medical image classification is a lack of sufficient annotated data for training deep networks from scratch. Fine-tuning existing CNNs that have been trained using a large annotated dataset from a different application is the best alternative to full training for medical applications. A second extremely interesting result it that training CNNs optimising a weighted ratio of two loss functions for simultaneous classification and regression provides a better error signal to the network and improves the overall classification performance. Many diseases are progressive by nature such as Alzheimer's disease, cancer, emphysema, tumours, lesions, and muscular dystrophy. Automatic quantification of such diseases using jointly trained CNNs may improve the quantification performance and provide insights to study the progression of the disease. Finally, it is very interesting that using multi-objective convolutional learning to jointly train CNNs based on different diagnostic features of a disease as ground truth can produce an overall improvement in the quantification performance achieving results on par with human accuracy. Multi-objective learning and joint prediction of multiple regression and classification variables can be useful to assess diseases involving multiple diagnostic features like Alzheimer's, multiple sclerosis, and multiple myeloma (cancer).

Bibliography

[1] H Oka, S Muraki, T Akune, A Mabuchi, T Suzuki, H Yoshida, S Yamamoto, K Nakamura, N Yoshimura, and H Kawaguchi. Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage*, 16(11):1300–1306, 2008.

[2] Lior Shamir, Shari M Ling, William Scott, Marc Hochberg, Luigi Ferrucci, and Ilya G Goldberg. Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage*, 17(10):1307–1312, 2009.

[3] Anne CA Marijnissen, Koen L Vincken, Petra AJM Vos, DBF Saris, MA Viergever, JWJ Bijlsma, LW Bartels, and FPJG Lafeber. Knee images digital analysis (kida): a novel method to quantify individual radiographic features of knee osteoarthritis in detail. *Osteoarthritis and Cartilage*, 16(2):234–243, 2008.

[4] Hee-Jin Park, Sam Soo Kim, So-Yeon Lee, Noh-Hyuck Park, Ji-Yeon Park, Yoon-Jung Choi, and Hyun-Jun Jeon. A practical MRI grading system for osteoarthritis of the knee: association with Kellgren–Lawrence radiographic scores. *European journal of radiology*, 82(1):112–117, 2013.

[5] David T Felson, Timothy E McAlindon, Jennifer J Anderson, Barbara W Weissman, Piran Aliabadi, Stephen Evans, Daniel Levy, and Michael P LaValley. Defining radiographic osteoarthritis for the whole knee. *Osteoarthritis and Cartilage*, 5(4):241–250, 1997.

[6] Hillary J Braun and Garry E Gold. Diagnosis of osteoarthritis: Imaging. *Bone*, 51(2):278–288, 2012.

[7] Lior Shamir, Shari M Ling, William W Scott Jr, Angelo Bos, Nikita Orlov, Tomasz J Macura, D Mark Eckley, Luigi Ferrucci, and Ilya G Goldberg. Knee X-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2), 2009.

[8] PG Conaghan, D Felson, G Gold, Stefan Lohmander, S Totterman, and R Altman. Mri and non-cartilaginous structures in knee osteoarthritis. *Osteoarthritis and cartilage*, 14:87–94, 2006.

[9] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis. *Source code for biology and medicine*, 3(1):13, 2008.

[10] Allan C. Gelber. Osteoarthritis research: current state of evidence. *Current Opinion in Rheumatology*, 27(3):273—-275, 2015.

[11] D Culliford, J Maskell, A Judge, C Cooper, D Prieto-Alhambra, NK Arden, COASt Study Group, et al. Future projections of total hip and knee arthroplasty in the uk: results from the uk clinical practice research datalink. *Osteoarthritis and Cartilage*, 23(4):594–600, 2015.

[12] Parastu S Emrani, Jeffrey N Katz, Courtenay L Kessler, William M Reichmann, Elizabeth A Wright, Timothy E McAlindon, and Elena Losina. Joint space narrowing and Kellgren–Lawrence progression in knee osteoarthritis: an analytic literature synthesis. *Osteoarthritis and Cartilage*, 16(8):873–882, 2008.

[13] DJ Hart and TD Spector. Kellgren & lawrence grade 1 osteophytes in the knee—-doubtful or definite? *Osteoarthritis and cartilage*, 11(2):149–150, 2003.

[14] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D Mark Eckley, and Ilya G Goldberg. WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern recognition letters*, 29(11):1684–1693, 2008.

[15] Jessie Thomson, Terence O'Neill, David Felson, and Tim Cootes. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*

[16] Honglak Lee. *Unsupervised feature learning via sparse hierarchical representations.* PhD thesis, Stanford University, 2010.

[17] Quoc V. Le. *Scalable feature learning.* PhD Thesis, Stanford University, 2013.

[18] Shulin Yang. *Feature Engineering in Fine-Grained Image Classification.* PhD Thesis, University of Washington, 2013.

[19] Claire Rebecca Donoghue. *Analysis of MRI for Knee Osteoarthritis using Machine Learning.* PhD Thesis, Imperial College London, 2013.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Osteoarthritis and Cartilage*, 42(1):60–88, 2017.

[21] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, (0), 2017.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[24] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013.

[25] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[26] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

[27] Holger R Roth, Amal Farag, Le Lu, Evrim B Turkbey, and Ronald M Summers. Deep convolutional networks for pancreas segmentation in ct imaging. In *SPIE Medical Imaging*, pages 94131G–94131G, 2015.

[28] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[29] M Subramoniam and V Rajini. Local binary pattern approach to the classification of osteoarthritis in knee x-ray images. *Asian Journal of Scientific Research*, 6(4):805, 2013.

[30] Barani Subramoniam et al. A non-invasive computer aided diagnosis of osteoarthritis from digital x-ray images. *Biomedical Research*, 2015.

[31] Dipali D Deokar and Chandrasekhar G Patil. Effective feature extraction based automatic knee osteoarthritis detection and classification using neural network. *International Journal of Engineering and Techniques*, 1(3), 2015.

[32] Tae Keun Yoo, Deok Won Kim, Soo Beom Choi, and Jee Soo Park. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: A cross-sectional study. *PloS one*, 11(2):e0148724, 2016.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[34] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya Goldberg. Wnd-charm: Multi-purpose image classifier. *Astrophysics Source Code Library*, 2013.

[35] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016.

[36] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35, 2017.

[37] Dieuwke Schiphof, Maarten Boers, and Sita MA Bierma-Zeinstra. Differences in descriptions of kellgren and lawrence grades of knee osteoarthritis. *Annals of the rheumatic diseases*, 67(7):1034–1036, 2008.

[38] Michael P LaValley, Timothy E McAlindon, Christine E Chaisson, Daniel Levy, and David T Felson. The validity of different definitions of radiographic worsening for longitudinal studies of knee osteoarthritis. *Journal of clinical epidemiology*, 54(1):30–39, 2001.

[39] Lisa Sheehy and T Derek V Cooke. Radiographic assessment of leg alignment and grading of knee osteoarthritis: A. 2015.

[40] Daniel L Riddle, William A Jiranek, and Jason R Hull. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons. *Orthopedics*, 36(1):e25–e32, 2013.

[41] Kristina Klara, Jamie E Collins, Ellen Gurary, Scott A Elman, Derek S Stenquist, Elena Losina, and Jeffrey N Katz. Reliability and accuracy of cross-sectional radiographic assessment of severe knee osteoarthritis: role of training and experience. *The Journal of rheumatology*, pages jrheum–151300, 2016.

[42] Lior Shamir, David T Felson, Luigi Ferrucci, and Ilya G Goldberg. Assessment of osteoarthritis initiative–kellgren and lawrence scoring projects quality using computer analysis. *Journal of Musculoskeletal Research*, 13(04):197–201, 2010.

[43] Y Sun, EC Teo, and QH Zhang. Discussions of knee joint segmentation. In *Biomedical and Pharmaceutical Engineering, 2006. ICBPE 2006. International Conference on*. IEEE, 2006.

[44] H Shaikh, Joshi Panbude, and Anuradha Joshi. Image segmentation techniques and its applications for knee joints: A survey. *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, 9(5):23–28, 2014.

[45] Shivanand S Gornale, Pooja U Patravali, and Ramesh R Manza. Detection of osteoarthritis using knee x-ray image analyses: A machine vision based approach. *International Journal of Computer Applications*, 145(1), 2016.

[46] Aleksei Tiulpin, Jerome Thevenot, Esa Rahtu, and Simo Saarakkala. A novel method for automatic localization of joint area on knee plain radiographs. In *Scandinavian Conference on Image Analysis*, pages 290–301. Springer, 2017.

[47] Tobias Stammberger, Felix Eckstein, Markus Michaelis, Karl-Hans Englmeier, and Maximilian Reiser. Interobserver reproducibility of quantitative cartilage measurements: comparison of b-spline snakes and manual segmentation. *Magnetic resonance imaging*, 17(7):1033–1042, 1999.

[48] Zohara A Cohen, Denise M Mccarthy, S Daniel Kwak, Perrine Legrand, Fabian Fogarasi, Edward J Ciaccio, and Gerard A Ateshian. Knee cartilage topography, thickness, and contact areas from mri: in-vitro calibration and in-vivo measurements. *Osteoarthritis and cartilage*, 7(1):95–109, 1999.

[49] Jukka Hirvasniemi, J Thevenot, V Immonen, T Liikavainio, P Pulkkinen, T Jämsä, J Arokoski, and S Saarakkala. Quantification of differences in bone texture from plain radiographs in knees with and without osteoarthritis. *Osteoarthritis and cartilage*, 22(10):1724–1731, 2014.

[50] Tomasz Woloszynski, Pawel Podsiadlo, GW Stachowiak, and M Kurzynski. A signature dissimilarity measure for trabecular bone texture in knee radiographs. *Medical physics*, 37(5):2030–2042, 2010.

[51] Feng Zhao and Xianghua Xie. An overview of interactive medical image segmentation. *Annals of the BMVA*, 2013(7):1–22, 2013.

[52] Cristian Dan Pirnog. Articular cartilage segmentation and tracking in sequential mr images of the knee. 2005.

[53] J Duryea, J Li, CG Peterfy, C Gordon, and HK Genant. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. *Medical physics*, 27(3):580–591, 2000.

[54] P Podsiadlo, M Wolski, and GW Stachowiak. Automated selection of trabecular bone regions in knee radiographs. *Medical physics*, 35(5):1870–1883, 2008.

[55] Lilik Anifah, I Ketut Eddy Purnama, Moch Hariadi, and Mauridhi Hery Purnomo. Automatic segmentation of impaired joint space area for osteoarthritis knee on x-ray image using gabor filter based morphology process. *IPTEK The Journal for Technology and Science*, 22(3), 2011.

[56] Hung-Chun Lee, Jiann-Shu Lee, Mark Chii-Jeng Lin, Chia-Hsiang Wu, and Yung-Nien Sun. Automatic assessment of knee osteoarthritis parameters from two-dimensional x-ray image. In *Innovative Computing, Information and Control, 2006. ICICIC'06. First International Conference on*, volume 2, pages 673–676. IEEE, 2006.

[57] Li Deng and Dong Yu. *Deep Learning.* Now Publishers Incorporated, 2014.

[58] Li Deng. Three classes of deep learning architectures and their applications: a tutorial survey. *APSIPA transactions on signal and information processing*, 2012.

[59] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. *CS231n: Convolutional Neural Networks for Visual Recognition*, 2016. Available at `http://cs231n.github.io/convolutional-networks/`.

[60] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

[61] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 2015.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[65] Shulin Yang. *Feature Engineering in Fine-Grained Image Classification.* PhD Thesis, University of Washington, 2013.

[66] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011.

[67] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

[68] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[69] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[70] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.

[71] Terry Anderson. Towards a theory of online learning. *Theory and practice of online learning*, 2:109–119, 2004.

[72] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

[73] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 52–59. Springer, 2011.

[74] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[75] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

[76] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[77] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[78] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[79] Irvin Sobel. An isotropic 3× 3 image gradient operator. *Machine Vision for three-demensional Sciences*, 1990.

[80] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[81] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 3(6):610–621, 1973.

[82] Simona E Grigorescu, Nicolai Petkov, and Peter Kruizinga. Comparison of texture features based on gabor filters. *IEEE Transactions on Image processing*, 11(10):1160–1167, 2002.

[83] Michael Reed Teague. Image analysis via the general theory of moments. *JOSA*, 70(8):920–930, 1980.

[84] Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE journal of biomedical and health informatics*, 19(5):1627–1636, 2015.

[85] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[86] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2015.

[87] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z Papadakis, Adrien Depeursinge, Ronald M Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–6, 2016.

[88] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

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

[90] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*, 2017.

[91] Patrick Ferdinand Christ, Florian Ettlinger, Felix Grün, Mohamed Ezzeldin A Elshaera, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, et al. Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*, 2017.

[92] Johannes Kilian. Simple image analysis by moments. *OpenCV library documentation*, 2001.

[93] Marco Bressan, Christopher R Dance, Hervé Poirier, and Damián Arregui. Local contrast enhancement. In *Color Imaging: Processing, Hardcopy, and Applications*, page 64930Y, 2007.

[94] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[95] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014.

[96] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[97] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of British Machine Vision Conference*, 2014.

[98] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

[99] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[100] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[101] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[102] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9:1871–1874, 2008.

[103] Shun Miao, Z Jane Wang, Yefeng Zheng, and Rui Liao. Real-time 2d/3d registration via cnn regression. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1430–1434. IEEE, 2016.

[104] Concetto Spampinato, Simone Palazzo, Daniela Giordano, Marco Aldinucci, and Rosalia Leonardi. Deep learning for automated skeletal bone age assessment in x-ray images. *Medical image analysis*, 36:41–51, 2017.

[105] Holger R Roth, Yinong Wang, Jianhua Yao, Le Lu, Joseph E Burns, and Ronald M Summers. Deep convolutional networks for automated detection of

posterior-element fractures on spine ct. In *Proceedings Volume 9785, Medical Imaging 2016: Computer-Aided Diagnosis*. SPIE Medical Imaging, 2016.

[106] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3459, 2015.

[107] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[108] René Ranftl and Thomas Pock. A deep variational model for image segmentation. In *German Conference on Pattern Recognition*, pages 107–118. Springer, 2014.

[109] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing*, 14(9):1360–1371, 2005.

[110] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[111] Christopher Beckham and Christopher Pal. A simple squared-error reformulation for ordinal classification. *arXiv preprint arXiv:1612.00775*, 2016.

[112] DT Felson, DR Gale, M Elon Gale, J Niu, DJ Hunter, J Goggins, and MP Lavalley. Osteophytes and progression of knee osteoarthritis. *Rheumatology*, 44(1):100–104, 2004.

[113] Junting Pan and Xavier Giró-i Nieto. End-to-end convolutional network for saliency prediction. *arXiv preprint arXiv:1507.01422*, 2015.

[114] Li Wan, David Eigen, and Rob Fergus. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 851–859, 2015.

[115] Basura Fernando and Stephen Gould. Learning end-to-end video classification with rank-pooling. In *International Conference on Machine Learning*, pages 1187–1196, 2016.

[116] Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4845–4849. IEEE, 2017.

[117] David A Williams, Michael J Farrell, Jill Cunningham, Richard H Gracely, Kirsten Ambrose, Thomas Cupps, Niveditha Mohan, and Daniel J Clauw. Knee pain and radiographic osteoarthritis interact in the prediction of levels of self-reported disability. *Arthritis Care & Research*, 51(4):558–561, 2004.

[118] Tim D Spector and Cyrus Cooper. Radiographic assessment of osteoarthritis in population studies: whither kellgren and lawrence? *Osteoarthritis and Cartilage*, 1(4):203–206, 1993.

[119] Simon K Warfield, Michael Kaus, Ferenc A Jolesz, and Ron Kikinis. Adaptive, template moderated, spatially varying statistical classification. *Medical image analysis*, 4(1):43–55, 2000.

[120] Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li, and Dimitris Metaxas. Automated anatomical landmark detection on distal femur surface using convolutional neural network. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 17–21. IEEE, 2015.

[121] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[122] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2016.