# Deep Learning-based Knee Osteoarthritis Diagnosis from Radiographs and Magnetic Resonance Images

Jiao Jiao
Matriculation number: 362187

A dissertation submitted in partial fulfillment of the requirements
for the degree of Master of Science in Media Informatics
at RWTH Aachen University, Germany

Supervisors:
Prof. Dr. Stefan Decker
Prof. Dr. med. Dr. rer. nat. Danilo Bzdok

Advisors:
Oya Deniz Beyan, Ph.D.
Michael Cochez, Ph.D.
Md. Rezaul Karim, M.Sc.

June 2019

# Eidesstattliche Versicherung

____________________________ ____________________________

Name, Vorname Matrikelnummer

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel

____________________________________________________________

____________________________________________________________

____________________________________________________________

selbständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

____________________________ ____________________________

Ort, Datum Unterschrift

*Nichtzutreffendes bitte streichen

**Belehrung:**

**§ 156 StGB: Falsche Versicherung an Eides Statt**

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

**§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt**

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

____________________________ ____________________________

Ort, Datum Unterschrift

# Contents

# List of Figures

# List of Tables

# Abstract

**Background:** with the increasing prevalence of knee Osteoarthritis (OA), a degenerative joint disease without effective and economical cure at the mid- to late- stage, the growing demand of its early diagnosis exposes an enormous challenge for highly precise, objective and efficient assessments. In terms of our investigation on deep learning based OA quantification mechanisms, this thesis presents a new approach to automate the OA diagnosis, which is based on multimodality integration concept.

**Methods:** our pipeline consists of data collection, preprocessing, Region of Interest (ROI) detection, classification and regression analysis, decision visualization, and ensemble module. In our approach, plain radiographs from coronal and sagittal plane are enhanced by using histogram equalization and slight Perona-Malik filter ($K$=50, iterations=1), while Magnetic Resonance Imaging (MRI) slices from sagittal and axial view are processed using slight Perona-Malik filter, unsharp masking sharpening with EDGE_ENHANCE kernel, and multi-slice averaging.

ROIs are then precisely extracted from the preprocessed images by means of U-Net with the ResNet-18 backbone, in which classification and regression analysis are carried out by DenseNet-161 and revised VGG-19, which are followed by highlighting the relative class-discriminating features using Gradient-weighted Class Activation Maps ++ (Grad-CAM++). Based on 2,406 individuals of the Multicenter Osteoarthritis Study (MOST) cohorts, existing technical bottlenecks such as noise, artefact, and modality limitation have been mitigated by the grading accuracy with experts of integrating Dense Convolutional Networks (DenseNets) and VGG-19 trained on coronal radiographs and sagittal MRIs.

**Results:** precision, recall, and the F1 scores of our OA grading approach are boosted to 0.90 by the combination of model ensemble (DenseNet-161, DenseNet-169, DenseNet-201 and VGG-19) on coronal X-Ray images and VGG-19 attained from sagittal MRIs. Moreover, ensemble experiments on coronal view of X-Ray images confirm that even if only assembling DenseNet-161, DenseNet-201 and VGG-19, it also served as an optimal alternative showing superior diagnosis accuracy of 91.22% with a reasonable Mean Squared Error (MSE) of 0.0878.

# Acknowledgements

Foremost, I intend to express my sincere gratitude to my supervisor, Prof. Dr. Stefan Decker, for the insightful comments and encouraging remarks addressed during my each presentation. And the Meetup *Machine Learning in Neuroimaging* given by my second supervisor Junior-Prof. Dr. med. Dr. rer. nat. Danilo Bzdok also inspired me throughout the entire technique selection and parameter tuning process of my master thesis.

In the meanwhile, my heartfelt thanks endlessly go to my advisors: Dr. Oya Beyan, Dr. Michael Cochez, and Mr. Md. Rezaul Karim. The door to their offices were always open whenever I ran into a trouble spot. They consistently allowed this paper to be my own work, but steered me in the right direction whenever I needed it. In order to guarantee I could focus more on technical tasks, Michael Cochez provided the continuous support of all the documentation work and RWTH Compute Cluster settings with patience and enthusiasm. In particular, I am gratefully indebted to his prompt and thorough feedback, corrections and valuable suggestions on the first draft of this thesis.

As for Md. Rezaul Karim, without his passionate participation and immense knowledge in the realm of deep learning, the issues, such as data acquirement application, complex label explanation, huge imbalance in the datasets and so on, could not have been successfully settled. He even enlightened me the first glance of machine learning research. And he also put considerable efforts on the final thesis correction.

Certainly, I also appreciate the swift assistance of Reinhard Linde, Claudia Puhl and Daniele Glöckner. Reinhard Linde established the connection for me to our GPU computing environment, which substantially boosted the progress of this project. And the whole process of thesis registration and submission was consummately taken care of by Claudia Puhl and Daniele Glöckner.

In the end, I am obliged to get access to plenty of clinical knee images and their assessments offered by MOST Coordinating Center, which laid the foundation of this topic.

# Chapter 1

# Introduction

As a phenomenal chronic disease, OA is defined as a heterogeneous group of conditions that lead to joint symptoms and signs that are associated with defective integrity of articular cartilage, in addition to related changes in the underlying bone at the joint margins [1]. The gradual loss of articular cartilage characterizes defective integrity. Videlicet, aging cartilage becomes stiffer progressively and more vulnerable to wear and tear [2].

## 1.1 Background and motivation

Nowadays, the increasing prevalence of OA not only has posed an enormous challenge for elderly health, but also tends to younger people with risk factors such as obesity and reduced physical activity [3]. As of 2006, about 40% of the population over the age of 65 have radiographic evidence of OA, in particular for knee, which is still dramatically soaring [4]. In terms of the studies by World Health Organization, the total number of Years Lived with Disability (YLDs) world-wide caused by OA increased by 60.2% between 1990 and 2010, and by 26.2% per 1000 individuals [3].

Ranked as the $13^{th}$ highest global disability factor [5], OA places a high economic expense burden on the society, in consideration of work absenteeism, early retirement, and joint replacement [6]. For example, as early as 1994, the cost of OA was estimated at $ 5.9 billion in Canada, which was mainly attributed to lost productivity at work and home due to disability [7]. When it comes to 2007, around $ 185.5 billion are estimated to be paid in United States as the aggregate annual expenditure [3]. Although with the formal inauguration of the novelty techniques, the cost of OA treatments began to shrink, up to 2015, the estimated expenditure per patient for joint replacement still reaches 19,715 €/year [3], which, to a great extent, arises from the current clinical inability to systematically diagnose the disease at an early stage [8]. After all, the incremental cost per patient with mild OA is calculated as 1300 €/year, while 4377 €is spent for severe patients yearly [3].

However, except expensive joint replacement surgery, there is no effective cure for OA at the mid- to late- stage [8]. Namely, early diagnosis remains the only available option to prolong patient health in a low-cost way.

Clinically, OA is primarily identified via hyaline cartilage change based on medical images. However, the process of measuring cartilage morphology is extremely time-consuming and burdensome, especially for MRI modality. Each 3-Dimensional (3D) knee MRI sequence takes up around six hours for a surgeon to manually analyze [9], which still cannot guarantee its diagnosis accuracy, due to the lack of additional information. Surgeons can hardly catch all the information presented in medical images, let alone stored metadata. Hence, supplementing the diagnostic chain with computer aid, radiologists and other clinical specialists can focus on incidental findings, instead of routine tasks like image grading. In the meanwhile, on the basis of accurate algorithms, the OA diagnosis can be conducted more comprehensive.

## 1.2 Problem statement

Early detection of OA is in the middle of a major paradigm shift driven by information technology. In particular, the booming of digital medical images is drawing more and more attention in the academic, even has engendered several novel branches in terms of modalities, including radiography (also named as X-Ray), MRI, Computed Tomography (CT) and ultrasound [2].

However, owing to different operating principles, each modality exposes several limitations, which make the accuracy of OA severity quantification and prediction meet the bottleneck fundamentally. For example, ultrasound requires so complex image processing algorithms that it is exceedingly costly both for processing time and for expense [1]. As for CT and radiograph, they exhibit significant potentials in displaying a bone shape rather than a soft-tissue region which provides a more explicit approach to vividly express characteristics of OA [10, 11]. Furthermore, CT is currently facing challenging technical dilemmas with respect to normalization strategies and controlling light beam angles[12]. X-Ray is stuck with subtle changes detection over time as well as in stereoscopic space. While MRI is greatly useful in identifying full or partial-thickness changes of articular cartilage, it cannot present clear bone architectures, which might indicate the earliest OA progression [2, 10, 13].

Apart from the aforementioned limitations of each modality, noises and artifacts, such as aliasing, slice overlap, truncation, and patient motion, acting as several pathological features of OA, i.e., osteophytes, bone marrow lesions and surface fibrillation, also restrict the flourish of automatic OA diagnosis so far [14, 15].

Thanks to the ground-breaking deep learning techniques, automatic medical image segmentation, a significant step for OA detection, has outperformed human experts [8]. Hence, currently, the core of research in automatic OA severity quantification realm is how to break through above three technical restraints so that it still possesses a promising potential to explore the dramatic performance enhancement.

## 1.3 Objectives and scope of the thsis

OA can result in any of joint in the body, such as spines, hips, fingers, etc, but it is most common and severe in knee [2]. Therefore, this thesis concentrates on knee OA analysis and prediction based on deep learning algorithms, on account of its remarkable prosperity in medical image segmentation field.

### 1.3.1 Thesis goal

Considering intrinsic modality restriction, noise and artifact would give rise to severe consequences, this thesis aims to sort out the ideal pipeline settings consisting of preprocessing, ROI extraction, classification, and multimodality integration.

Specifically, for the sake of noise removal as well as artifact reduction, this thesis proposes to figure out the most suitable group of preprocessing operations. Furthermore, in order to promote the classification accuracy, our research intends to investigate the perfect deep learning based ROI extraction algorithms for medical image. Moreover, targeting at resolving intrinsic restrictions for each mainstream modality, this thesis is designed to validate the superiority of multimodality integration and further explore the most fitting modality combination on MOST public database, one of the standard datasets applied in knee OA researches.

### 1.3.2 Scope of the thesis

In the present study, a novel OA quantification solution based on multimodality integration is applied to overcome the limitations of the state-of-the-art approaches by effectively getting rid of the negative influences stemming from modality-forming principles. In terms of their complementarity and application range, following a series of preprocessing methods, including contrast enhancement, noise elimination, and multi-slice integration, aiming at the same patient, radiographs and MRIs from axial, sagittal as well as coronal plane are respectively classified by their extracted features in the detected ROIs. Given supplying reliability and objectivity of the grading process, class-discriminating attention maps are generated by Class Activation Maps (CAM), prior to the averaging ensemble of models with the same modality and multimodality.

### 1.3.3 Key contributions

The key contributions of this thesis can be summarized in the following:

1. Aiming at plain radiograph and MRI, found the best preprocessing approach respectively based on histogram equalization, Perona-Malik filter, unsharp masking edge enhancement and MRI slice average.

2. Developd an accurate framework to automatically detect and localize knee joints from different perspectives of X-ray images using a U-Net with ResNet-18 as backbone.

3. Taking advantage of their both natural benefits, presented a new approach to train a ROI detection network on plain radiographs with the combination of two loss functions: Binary Cross Entropy (BCE) and Intersection-over-Union (IoU).

4. Validated the performance of Region Proposal Network (RPN) in Faster Region Convolutional Neural Network (FRCNN) for ROI detection and came up with the shortcomings of end-to-end learning for knee OA grading.

5. Proposed a highly precise architecture, fully convolutional VGG-19, to alleviate overfitting and resolve the excessive parameter issue of fully connected layers in VGG-19, in particular for medium and large sized medical images.

6. Refined DenseNet-161 as the remarkably accurate OA predictor assessed on various views of both X-Ray images and MRIs.

7. Validated the promotion of diagnosis performance dramatically by introducing dataset balancing based on image augmentation.

8. Compared to Kellgren and Lawrence (KL) semi-quantitative metric, validated the superiority of Osteoarthritis Research Society International (OARSI) Joint Space Narrowing (JSN) progression scoring system on the basis of empirical outcomes.

9. Brought in Grad-CAM++ as OA diagnosis visualization to more precisely locate class-discriminating features.

10. Framed model ensemble concept on diverse neural networks and substantiated its better performance than assembling the same series of models.

11. Introduced multimodality integration conception to break through limitations of plain radiograph and MRI themselves, whose promising boost and prospect were confirmed by integrating models obtained from coronal X-Ray images and sagittal MRIs.

The rest of the thesis is structured as follows: chapter 2 gives a panoramic view of the related works on knee OA quantitative criteria as well as its automatic diagnosis methodologies and outlines their potential limitations. In terms of papers analyzed in chapter 2, chapter 3 explains the desirable approach and relative algorithm candidates for further evaluation. chapter 4 (Evaluation) provides the experimental comparative performance for the proposed mechanism in chapter 3 and sorts out the ideal implementations for each module. Finally, chapter 5 summarizes the accomplishments for the work and points out future directions which might lead towards the next generation of systems.

# Chapter 2

# Related Work

Owing to the ever-growing prosperity of medical image modalities, instead of aiding with a differential diagnosis where there is clinical doubt, imaging plays a pivotal role in the OA diagnosis and prediction of structural progression. In particular, modalities have a predominant influence on determination as well as long-term prognosis of OA. Namely, diverse patterns contribute to a different understanding of OA pathogenesis. Therefore, a panorama of automatic OA quantization criteria and mechanisms is proposed based on the analysis of employed image modalities, including radiography, MRI, CT, ultrasound, and so on.

## 2.1 Radiographs

While different modalities have expanded our noesis on OA pathologies by offer embracing 3D perspectives on the joint, conventional radiographs are still the first line imaging modality selected, on account of its availability, swiftness, and economy [16]. A radiographic image is a shadow of the differential absorption of x-rays by the tissues of the joint, where bony structures appear white to light grey whereas the radio-transparent soft tissues turn out dark grey to black [17]. Radiographs exhibit significant potentials in displaying bone shape features related to OA, inclusive of osteophytes, Joint Space Width (JSW) narrowing, subchondral sclerosis and cysts, rather than adjacent tissues which provide a more explicit approach to vividly express characteristics of OA, in particular for the early detection [10, 11, 18, 19]. After all, those pathological changes shown in radiographs are associated with severe stages of OA [18, 19]. Even among experts, the first OA degree reported from radiological and clinical diagnosis may be divergent [20].

Furthermore, due to the projection nature, radiography is stuck with subtle changes detection longitudinally as well as in stereoscopic space. Certainly, to a great extent, the more perspectives are examined, the higher the likelihood of correct knee OA diagnosis is [20]. Besides, exposure to ionizing radiation has already aroused the public worry about patient safety [18, 19]. Although radiography is not like the gold standard to detect and monitor osteoarthritis progression anymore, its widespread application in the evaluation of OA progression still fosters the maturity and variety of related studies, especially in the fields of admitted scoring systems and quantification workflows.

### 2.1.1 Evaluation criteria

Albeit OA is radiographically defined by the presence of marginal osteophytes, worsening of JSN serves as the most common indication of progressive osteoarthritis, which is assessed via quantitative JSW or semi-quantitative grading systems [21].

#### Quantitative assessments

JSW is the distance between the projected femoral and tibial margins on the anteroposterior radiographic image [21]. In general, minimum JSW is the standard metric. Furthermore, measures of location-specific JSW with various degrees of responsiveness is substantiated to be superior to minimum JSW assessment concerning the prediction of progression [22]. Nevertheless, both JSW metrics are reliable when obtained radiographs last longer than two years, and knees are fixed in a standardized flexed position since their minor changes are in the millimeter/submillimeter range, which is high-demanding for precision [21, 22].

Investigations of quantitative evaluation are not limited to joint space. Fractal Signature Analysis (FSA) of the subchondral bone has been a matter of debate in OA research [22]. FSA extracts the trabecular texture of subchondral bones in horizontal and vertical vectors and revealed correlations with the degree of cartilage loss on the tibial surface longitudinally [22]. However, instead of standard digital radiography for our routine OA examination, barely macro radiography (typically with 4-5 times magnification) allows for better visualization of structural changes in FSA calculations [22]. Noticeably, FSA still calls for massive technical renovation in clinical practice.

#### Semi-quantitative assessments

Due to in difficulty of measurement interpretation, semi-quantitative grading systems vastly transcend quantitative metrics in automatic knee OA detection, among which the KL scale and OARSI atlas retain dominant. KL scale defines radiographic OA with a global composite score on a 0-4 range [23], which is correlated to incremental severity of OA, with grade 0 signifying no presence of OA and grade 4 indicating grievous OA [24]. The following radiologic features are crucial factors for KL [24]:

- Formation of osteophytes on joint margin or the tibial spines
- Periarticular ossicles (basically about distal interphalangeal joints and posterior interphalangeal joints)
- Narrowing of joint cartilage associated with sclerosis of subchondral bone
- Small pseudocystic areas with sclerotic walls situated usually in the subchondral bone
- Altered shape of bone ends, particularly in the head of the femur

Accordingly, representative knee radiograph with each KL grade is demonstrated in Table 2.1 [23].

| Grade | Radiologic symptom |
|---|---|
| 0 | None |
| 1 | Doubtful narrowing of the joint space with possible osteophyte formation. |
| 2 | Possible narrowing of the joint space with definite osteophyte formation. |
| 3 | Definite narrowing of joint space, moderate osteophyte formation, some sclerosis and possible deformity of bony ends. |
| 4 | Large osteophyte formation, severe narrowing of the joint space with marked sclerosis and definite deformity of bone ends. |

**Table 2.1:** KL grade description

However, employing the KL system for progression description has entailed additional risks for automatic quantification, since KL grade 3 includes all degrees of JSN, regardless of the actual extent [21, 22]. Moreover, knees at KL grade 4 that exhibit a bone-on-bone appearance are still prone to structural changes, such as bone marrow lesions, effusion, synovitis, and Hoffa-synovitis, which only can be detected on MRI [21, 22]. Therefore, the term "end stage" is no longer appropriate for KL grade 4 and a modified KL definition engenders [25].

Assessment of individual radiographic features of OA has been advocated as an alternative to KL score. The OARSI atlas provides image examples for grades for specific features of OA rather than assigning global scores according to definitions [22]. This atlas grades tibiofemoral JSN and osteophytes separately for each compartment of knees (medial tibiofemoral, lateral tibiofemoral, and patellofemoral) with a 0-3 scale according to the following guidelines in Table 2.2 [21, 25, 26].

Even if the usage of this atlas can standardize the interpretation of radiographs in trials and has moderate to better reliability than KL systems, the inherent limitations on variations in beam angle or knee flexion are not eliminated [25, 27]. Thereby, grades assessed with JSN is more rigorous than that of osteophytes [27]. Considering the time-consuming

| Grade | Criteria description |
|---|---|
| 0 | None |
| 1 | Mild (1%-33% abnormal) |
| 2 | Moderate (34%-66% abnormal) |
| 3 | Severe (67%-100% abnormal) |

**Table 2.2:** OARSI atlas criteria description

rating of the OARSI atlas, typically only OARSI JSN scores are taken into use among the scholastic [27].

To sum up, among different radiographic metrics, by their evaluation convenience, KL grade sticks out as representatives to be chiefly publicized in automatic OA detection. However, with the escalation of available clinical data, the rising of OARSI JSN score in the computer-aid diagnosis domain commences, which contributes to its higher confidence level.

### 2.1.2 Quantification workflows

Thanks to the abundant data sources and long-term in-depth research, the automatic OA diagnosis workflow of radiograph is versatile and flexible, which primarily consists of preprocessing and classification.

#### Preprocessing

Under its plain image format, generally, radiograph preprocessing just horizontally flips images for data augmentation [8, 28]. Undoubtedly, aiming at specific datasets, data cleaning, such as excluding images with implants to avoid any disturbances in the data distribution, is also emphasized [8]. However, to a great extent, a radiograph is simple and clear enough for direct classification after data augmentation.

#### Classification

Initially, previous to classification, the typical process comprised of feature extraction step. For example, Tae Keun Yoo et al. calculated Kinematic factors for stair ascent as features, which are classified by Support Vector Machines (SVM) with 97.4% accuracy [29]. Certainly, only six patients' radiographs collected by themselves cannot convince the public of their overwhelming success. Having contradistinguished with naive Bayes, Radial Basis Function (RBF) networks and random forest, [30] offered empirical evidence of SVM's superexcellence, especially in the aspect of sensitivity and recall rate. In the meanwhile, the authors imported Restricted Boltzmann Machine (RBM) to select features on the images after wavelet transformation [30], which inspired the extensive exploration for the adhibition of deep learning techniques in knee OA analysis.

Foremost, the failure of unsupervised Artificial Neural Networks (ANN) has been confirmed. L. Anifah et al. applied Gabor kernel to extract features from normalized radiographs obtained from Osteoarthritis Initiative (OAI) [31]. However, their Self Organizing Map (SOM) model cannot produce a satisfying classification accuracy (lower than 30% overall), which is far away from the requirements in practice [31]. Although [2] accreted Gray-Level Co-Occurrence Matrix (GLCM) with Gabor kernel for ROI extraction, which elevated their detection accuracy to 53.34%, only 4% of radiographs correctly detected as KL grade 2 obviously cannot achieve public expectation, which proves at least SOM should be eliminated in the further OA research.

Despite their failures, the amelioration rooting in ROI abstraction attracts the academic to build up their framework with this step, such as another method proposed in [2]. M. Subramoniam and V. Rajini manually selected 200*200 central joint region as ROI, where further they extracted features by Local Binary Pattern (LBP) with several distance metrics [2]. Although the trained K Nearest Neighbor (KNN) classifier presented over 95% accuracy, just 50 samples are quite difficult to produce a convincing result. Besides, their ROI selection is not only exhausted but also lack of evident scientific support. Building on this concept, [32] picked up three small regions along the tibial edge from medial view to lateral view, which bore out the rationality of the middle choice by experimental comparative deviations.

Through the comparison among KNN with the Euclidean distance, SVM with the Radial Basis Function (RBF) kernels, the logistic regression with the binomial model and the naive Bayes classifier, SVM stood out again from the crowd [32]. The 80.38% accuracy was verified by multiple wavelet decomposed features of 688 radiographs from OAI database, nevertheless, different with other experiments, only KL grade 0 and 2 were involved in this binary classification, which apparently would lead to the precision surge. Similar operations with random forest classifier in [33] revealed this conclusion on the MOST dataset. For the binary classification task, the grades were split into two groups: non-OA: KL (0,1) and the OA group: KL (2-4), which reached 80% accuracy or so. Correspondingly, 5-Class classification maintains the accuracy of around 45%. Moreover, to shun the onerous ROI labeling, [33] put forward the Random Forest Regression Voting Constrained Local Model to locate more informative points in both single bones and combinations of bones, which would compose ROIs. Although at least 4% difference between manual annotation and automatic tagging, it enlightens the academia on the direction for fully automated OA quantization.

With independence from prior knowledge and human effort in shift-invariant feature design, Convolutional Neural Network (CNN) gained widespread attention gradually. After all, the connectivity pattern between neurons resembles the organization of the animal visual cortex, which automatically extracts salient features for human determina-

| Reference | Approach | Accuracy | Limitation |
|---|---|---|---|
| [2] | SOM | 53.34% | Poor classification accuracy, in particular for KL grade 2 |
| [2] | KNN | 97.37% | Not enough subjects (50 patients) involved into experiments |
| [8] | Siamese CNN | 66.71% | Dissatisfying classification accuracy, in particular for KL grade 1 |
| [28] | CNN | 61.9% | Dissatisfying classification accuracy |
| [29] | SVM | 97.4% | Too less subjects (6 patients) involved into experiments |
| [31] | SOM | 29.53% | Poor classification accuracy |
| [32] | SVM | 80.38% | Only binary classification for KL=0 and KL=2 |
| [33] | Random forest | 47.9% | Poor classification accuracy |
| [34] | CaffeNet | 59.6% | Dissatisfying classification accuracy and poor ROI detection accuracy |
| [35] | DenseNet | —— | Unrecognized classification criterion by KL grade combination |
| [36] | CNN | 68% | Massive metadata required and class combination between KL grade 0 and 1 |

**Table 2.3:** Conclusion of knee OA diagnosis related work based on radiograph

tion. Thus, the whole process integrates into ROI detection and classification. For example, Joseph Antony et al. applied a linear SVM with the Sobel horizontal image gradients to detect knee joints and fine-tuned the pre-trained BVLC CaffeNet and VGG-M-128 by OAI database [34]. As for each detected ROI, they evaluated by the well-known Jaccard index, or essentially IoU to give the following matching score in eq. (2.1):

$$J(X,Y) = \frac{X \cap Y}{X \cup Y} \tag{2.1}$$

Where X is the manually annotated and Y is the automatically extracted knee joint center, their mean Jaccard index is only 0.36 [34]. Although the proposed method with 59.6% accuracy also left much to be desired, its greatest contribution stems from bringing in regression analysis. The continuous quantization was certified to both reduce the mean squared error and enhance the multiclass classification accuracy of the model from 57.6% [34]. After they upgraded linear SVM to Fully Convolutional Neural Network (FCN), inspired by the success of FCN for semantic segmentation on common pictures, the mean Jaccard index sharply leaped to 83%, even if OAI and MOST datasets are combined [28].

However, having trained their own 6-layer CNN model, 61.9% classification accuracy, whose gap between automatic and manual localization narrowed down from 4% in [33] to current 2%, still didn't live up to clinical desire.

Aleksei Tiulpin et al. applied SVM-based scoring again for gradually modifying proposed ROIs [37]. On the extracted joint areas, a 6-layer deep Siamese CNN detected the lateral and medial sides of the knee joint in the two branches respectively [8]. Aroused by [38, 39, 40], they ensembled Grad-CAM for Siamese networks attention visualization, to automatically highlight the important features to produce the target label, which fortifies the reliability and transparency of the whole architecture. Compared with the fine-tuned ResNet-34 (Residual Network with 34 layers) pre-trained on the ImageNet dataset, its average multiclass accuracy (66.71%) was a bit less, although its MSE was lower. Even if [41] was also pre-trained by ImageNet, it (67.2%) cannot yet catch up with ResNet-34 (67.49%). However, it proposes the hypothesis based on FRCNN ushering in a new era for the end-to-end classification without intermediate steps.

In conclusion, currently, two modes of classification prevail in the academic: ROI extraction + classification and end-to-end classification. Nevertheless, as Table 2.3 indicated, no matter which architecture is applied, 68% is deemed insufficient for practical application. The imprecision springs from stage 1 and 2, especially the former, which has no striking difference with normal joints in insensitive radiographs. Although [35, 36] merged stage 0 and stage 1 to attain over 70% precision, forasmuch as the restriction of this actual condition, OA prediction accuracy is facing a huge bottleneck, which bare criterion or workflow promotion manifestly cannot surmount.

## 2.2 MRI

As the stagnation for the radiographic OA progressed, gradually the emphasis of scholars shifts to MRI, a non-invasive modality with high spatial resolution without ionizing radiation and multi-planar capability that allows direct visualization of all structures of the joint, peculiarly cartilage morphology and biochemical composition [42, 43].

Clinical MRI is based on the alignment behavior (typically to 90° with the commonly used spin-echo technique) of hydrogen nuclei in a magnetic field [44]. When resonating with the applied external radio frequency pulse, the radio frequency signal produced by excited hydrogen nuclei realignment constructs a detailed superior contrast 2D and 3D image [18, 44]. Owing to water as the major form of hydrogen in the knee region, MRI depicts bone tissue as a signal void, while cartilage, muscle, ligaments and other soft tissues exhibit with low signal intensity [45]. MRI is greatly sensitive in identifying full or partial-thickness changes of articular cartilage over time, on the contrary, it cannot present clear bone architectures, which is in contrast to the technical characteristics of radiography [2, 10, 13].

As the most complex modality, MRI struggle with its diversity comprised by varying image planes (Axial, Sagittal, Coronal), multifarious sequences (Spin-Echo, Fast Spine-Echo, Gradient Echo and Inversion Recovery) and miscellaneous parameters (Proton density, Relaxation times, Magnetic field power), which strongly change MRI signals and require excessive robustness [14]. Thus, the crux of blooming MRI OA research lies in the substantial collection of standardized MRI, rather than regulated evaluation criteria and workflows.

### 2.2.1 Evaluation criteria

In comparison to the popularity of radiographic OA semi-quantitative assessment, quantitative metrics almost monopolize the academia of MRI OA diagnosis due to the high-resolution image sequences with the convoluted indices.

#### Quantitative assessments

The blossom of quantitative measurement of biochemical and biomechanical properties of the articular cartilage on MRIs has even compelled the formulation of quantitative MRI (qMRI), an advanced sub-modality in vivo [19]. qMRI for measuring the relaxation properties in cartilage, such as T2 mapping, T1 mapping, T1$\rho$ mapping, and T2* mapping, may aid in the diagnosis of early OA before irreversible morphologic changes [19, 46]. Those relaxation times characterize the fluid (water protons) or proteoglycan, which reflects as the signal intensity. For example, tissues with a strong interaction between hydrogen nuclei and the electromagnetic vibrations of macromolecules (fat tissue) exhibit a short T1 relaxation time and are bright on a T1-weighted image [44].

Similarly, abnormal symptoms caused by accumulated free water molecules (not attached to adjacent macromolecules) display long T2 relaxation times and bright signal intensity on T2-weighted images [44]. Apart from relaxation time, other MR parameters like proton density also can be associated with the change in water content and collagen fibril network for quantification [19]. However, growing investigations engage in a more intuitive metric, cartilage volume and thickness measurements, thanks to the super-resolution of MRI. Exactly, in this case, the accuracy of cartilage segmentation would face unprecedented challenges.

#### Semi-quantitative assessments

In the past decade, five well-established MRI scoring systems were published: the Whole-Organ Magnetic Resonance Imaging Score (WORMS), the Knee Osteoarthritis Scoring System (KOSS), the Boston Leeds Osteoarthritis Knee Score (BLOKS), the MRI Osteoarthritis Knee Score (MOAKS) and the Knee Inflammation MRI Scoring System (KIMRISS), among which WORMS and BLOKS have been broadly disseminated [22, 47].

Similar to OARSI atlas, both assessments approach highlighted above examine a spectrum of OA-related structural abnormalities including soft tissue, cartilage, and bone in the knee at various anatomical subregion locations [48]. Confronted with global indices, they offer highly reliable insights in cartilage morphology, bone marrow lesions, meniscal damage, and mediolateral meniscal extrusion, especially BLOKS, is more sensitive for full thickness defects [48]. Nevertheless, integration over ten articular surface regions concerning over ten independent features is too cumbersome for automatic diagnosis to handle. Hence, state-of-art MRI classifications predominantly stem from its quantitative metrics, even radiographic semi-quantitative metrics.

In comparison to the persistence of radiographic OA measurement standards, criteria of MRI assessments burst into bloom, which leads to the lack of authoritative labeled datasets so that radiographic metrics have surprisingly abounded in the academic.

### 2.2.2 Quantification workflows

The kernel of radiographic prediction is classification, whereas the established procedure for MRI processing highlights the preprocessing and segmentation step, contingent on the complication of image format itself.

#### Preprocessing

The superb multi-tissue assessment by MRI is derived from the sophisticated image modality, which brings about the indispensable standing for preprocessing. In particular, data acquired directly from the clinic are a series of 3D videos. Luckily, major acknowledged databases have projected them into 2D slices from diversified perspectives. Even though MRI is capable of imaging the soft tissues, improper contrast distribution, and brightness distribution always make for incorrect edges between the adjacent tissues [49]. Hence, those 2D MRI slices are subjected to contrast enhancement firstly. Certainly, linear rescaling can normalize MRIs to a fixed intensity range for brightness redistribution [50, 51], but it exerts a limited effect on dark medical images.

Histogram equalization, as the most common approach, adapts pixels to suitable grey level distribution, which generally augments white pixels for a better view of anatomical boundaries [11]. With the success of histogram equalization [11, 14, 52, 53, 54], a battery of variants have sprung up, such as Bi-histogram based Histogram Equalization, Hierarchical Correlation Histogram Analysis, Recursive Mean Separate Histogram Equalization, Recursive Sub-Image Histogram Equalization, Bi-histogram based Bezier Curve Contrast Enhancement and Spline-based Contrast Enhancement [49]. However, the techniques mentioned above only validated on 8 bits of images with grayscale values from 0 to 255. Grayscale for different MR pulse sequences varies from 0 to 1590. To exhibit the adequateness in the global environment, [49] developed Local Gray Level Transformation using S-curve technique, which efficiently represents enough contrast difference between tissues with maximized grey level increase and minimized decrease.

Following contrast enhancement, thresholding operation is then performed to exclude pixels whose intensity is less than half the average intensity of the image, which is appropriate for further binary image segmentation [11, 52, 54, 55].

As referred in chapter 1, noise and artifact block the development of high-precision OA quantization, in particular for MRI. Thus, their removers, filters, are imperative for the whole workflow. A. Paproki et al. employed the basic median smoothing with radius 1*1*1 [50]. To remove unnecessary high-frequency edges around cartilages, [41] picked on Gaussian low-pass filtering and [51] integrated sigmoid filter. These algorithms also cut off details by smoothing like [50] so that they are merely applicable for classifying the thickness of cartilages eventually.

After an empirical comparison by A. Suponenkovs et al., Perona-Malik filtering for gradient anisotropic diffusion smoothing stood out [14, 51, 53]. In addition to common noises, the bias of the magnetic field is the special artifact of MRI, which also can be dealt with filters, such as bias correction field filter in SimpleITK [51]. Moreover, as highlighted in chapter 1, MRI struggles with diversiform pulse sequences and parameters, where Sobel filter leveraged its power by calculating their derivatives [14].

Compared with the Sobel filter, affine registration is the riper way to solve the multiformity of pulse sequences and parameters. The obtained MRI affine transformation was propagated to the average surface for alignment, which integrates the intermediate result for next segmentation by MIRROR estimation algorithm [50, 53]. Broadly, affine registration aims to deal with MRI voxels instead of 2D MRI slices.

Hitherto, contrast enhancement, intensity regulation, noise elimination, and different measurement integration are emphasized in MRI preprocessing. Distinctly, those ideas also can be extended to other modalities, in particular for the primitive radiography.

### Segmentation

As the major approach for MRI ROI extraction, in general, segmentation can be divided into 2D and 3D model. According to the stored data format of existing public databases, this section concentrates on 2D segmentation. The maturity of radiographic segmentation and edge detection approaches still act as a leading position for MRI segmentation. Followed by masking, generated segmented cartilage (ROI) automatically, Canny edge detection determines edges by identifying local maxima of the image gradient [11, 52]. Undoubtedly, this idea is only befitting for cartilage thickness classifier in the next step. [55] framed a similar concept that ROIs were extracted via the largest blob detection, which binarized images after a certain intensity threshold in preprocessing operation to form a convex image mask for meniscus. Following the blossom of Machine Learning (ML), MRI segmentation based on classification gains steam. These methods classify each

| Reference | Approach | Limitation |
|---|---|---|
| [9] | ANN | Not enough subjects (100 patients) involved into experiments |
| [14] | K-means | Only statistical analysis of selected features |
| [52] | SVM | Too less subjects (15 patients) involved into experiments |
| [54] | SVM | Too less subjects (16 ROIs) involved into experiments and only binary classification for anomaly detection |
| [56] | —— | Only statistical analysis of T2 |
| [57] | —— | Only statistical analysis of T1 |
| [58] | GHMM | Diseased region detection focus |

**Table 2.4:** Conclusion of knee OA diagnosis related work based on MRI

voxel into two classes, ROI and the other. For example, Dong Yang et al. sliced 3D MRIs into three sets of 2D images with X, Y, Z axes, respectively. For each selected landmark, 2D images are labeled as either positive or negative for the corresponding CNN classifier based on whether they contain this landmark [59]. With the assistance of Procrustes analysis, those predicted landmarks could successfully calculate parameters of rigid transformation for the average training mesh (initial boundaries), which fulfilled the femur segmentation. Although this scheme has accomplished certain constituent segmentation preeminently, different components in knee MRIs possess individual standings in OA diagnosis, which cannot unitize as a global ROI. Hence, [60] introduced a brand new framework that defines multiple classifiers dependent on location, Location-Dependent Image Classification (LDIC).

In brief, it decomposed the whole image into a set of cells, whose intensity as features for further Gaussian Mixture Model (GMM) based classification. Through Genetic Algorithm (GA), those classified cells are iteratively grouped by a heuristic search. This idea only produced better performance when the appropriate combination of cells was set as GA's initial individuals. A. Suponenkovs et al. adopted k-means clustering to achieve above formulation since it is possible to control the number of clusters in case of creating so many segments like Watershed algorithm [14]. Nevertheless, the initial centroid selection faced the same dilemma with [60].

Consultative hints from recent 3D segmentation models have emerged, such as FCN established by [61] and U-Net employed in [62], whose encoder-decoder structure was specially designed for biomedical image segmentation. Uncannily paralleling with U-Net architecture, Generative Adversarial Network (GAN) as well as its variants, like conditional Generative Adversarial Network (cGAN), have been resoundingly demonstrated in other MRI segmentation, where 2D models can draw brilliant inspirations [63, 64].

### Classification

Contrary to the ripeness of radiographic OA quantization, relative MRI research is scarce, since the plurality of papers stalls at the segmentation process. After all, having witnessed to the proof by [56, 57, 65] that T1 and T2 relaxation time can characterize articular cartilage tissue, the academic can readily acquire OA severity from T1 or T2 calculated by the segmented image intensity. However, articular cartilage is not the only determinative factor for OA. T1 or T2 mapping may provide valuable information on cartilage morphological and biochemical changes, which gives more convictive evidence for assessment of knee OA progress, rather than a true gold criterion [56, 57].

Additionally, considering the core operation, intensity thresholding, calls for the intensity disparity, the application of T1 and T2 is technically demanding, which merely support formally standardized but enhanced MR scans with high-precise segmentation [57]. Thereby, A. Suponenkovs et al. barely referred to the trend between healthy subjects and patients figured out from the dispersion method (T2) [14]. [54] emphasized Stereological and Textural Measurements (STM), which were detrended by the General Linear Model (GLM). Thanks to Principal Component Analysis (PCA) of normalized STMs, top 5% informative features can be selected efficiently. However, in the case that their SVM classifier with an RBF Gaussian kernel just carried on anomaly detection (binary classification), maximum 73% accuracy among 16 ROIs doubtless disappointed the public.

S. Kubakaddi et al. attempted to quantify segmented cartilage thickness [52], which implemented in [11] by calculated ROI GLCM features. Although 86.66% SVM accuracy based on 15 patients still has to be verified with larger subtype sample sizes, at least they have been confirmed the enhancement from STM features. Aware of the authority of KL and its convenient access, Chao Huang et al. came up quantification with Gaussian Hidden Markov Model (GHMM) whose parameters are estimated by Expectation-Maximization (EM) algorithm [58]. It designed for both diseased region detection in each OA subject and localized analysis of longitudinal cartilage thickness within each latent subpopulation [58].

However, void of verification by recognized datasets, they even didn't publish their classification conclusion. Likewise, [9] also predicted KL scores for MRI, but it focused on the comparison among ML algorithms, including ANN, SVM, random forest, and naive Bayes. Based on PCA for all the 36 informative locations, ANN distinguished from others with 0.714 F-Measure [9]. Although 100 pairs of knees were lack of credibility, owing to the qualitative leap of quantification performance, Y. Du et al. still demonstrated the potential of MRI, KL, and ANN in the realm of OA diagnosis.

In brief, behind the trend towards end-to-end learning for radiographic OA analysis, MRI OA detection explicitly is divided into preprocessing, segmentation and classification. Although masses of studies stall at the segmentation for its modality complexity, as Table 2.4 shown, they still have affirmed the promising combination of MRI, deep learning and radiographic OA metrics.

## 2.3 CT

Originally, CT is primarily applied for the brain and lung with connective tissue [16]. The feasibility as an arthroscopic based clinical instrument and the ability to assess joint cartilage on a micron scale have initiated the emergence of CT in OA research [66]. It generally detects bone abnormalities in the axial skeleton or other joints where radiographs are unclear, and MRI is contraindicated, especially for hip OA [16].

With superior images of the bony cortex and soft-tissue calcification, CT should serve as a reasonable gold standard in OA research when validating bone morphology such as cysts, erosions, and osteophytes [16]. Nevertheless, its two leading limitations: low soft-tissue contrast and radiation are indeed more serious than that of other modalities so that CT hasn't independently fulfilled a knee OA quantization, not to mention an entire workflow or evaluation criterion.

Heretofore, Y. Uozumi et al. binarized the raw CT images and formed a bone region of the femur and tibia by masking, but those operations were just the preparation for further MRI segmentation [10]. Blending with advanced techniques, [67] and [68] exploited its progressive variant, Phase Contrast Imaging X-Ray Computed Tomography (PCI-CT), for OA quantitative characterization. Texture features in [67] derived from Minkowski Functionals (MF) and GLCM were classified by Support Vector Regression (SVR) with a radial basis function kernel.

As for [68], with the same GLCM features constructed in the designated ROIs, they directly evaluated OA severity via CaffeNet. However, none of them present an experiment with over five patients as a reference. [69] classified Cone Beam Computed Tomography (CBCT) scans, another variant of CT, by deep neural networks, which also faced the plight of accessible data shortage. Hence, those high Area Under the receiver operating characteristic Curves (AUC) demonstrated from above researches rooted in the repetition images from the same patient, which is overtly lack of science and rationality. Since academic is unable to attest the prominent advantage of CT till date and it is highly challenging to get access to a credible CT database, as shown in Table 2.5, CT OA diagnosis framework still leaves a massive gap between the current situation and practical demand.

| Reference | Approach | Limitation |
|---|---|---|
| [67] | SVR | Too less subjects (5 patients) involved into experiments |
| [68] | CaffeNet | Too less subjects (5 patients) involved into experiments |
| [69] | ANN | Unrecognized classification criterion, web application focus and too less subjects (34 patients) involved into experiments |

**Table 2.5:** Conclusion of knee OA diagnosis related work based on CT

| Reference | Approach | Limitation |
|---|---|---|
| [70] | RW | Only segmentation involved |
| [71] | Histogram equalization | Only preprocessing involved |

**Table 2.6:** Conclusion of knee OA related work based on ultrasound

## 2.4 Ultrasound

The expensive MRI leads to the ultrasound as an alternative imaging tool for quantitative assessment of femoral cartilage thickness, since ultrasound possesses the properties: accessibility in conjunction with plain radiography, portability and allows for real-time image acquisition [22, 70]. Podlipska et al. corroborated that ultrasound examination is extra beneficial to the depiction of meniscal extrusion, even may be superior to plain radiography for changes in medial femoral cartilage morphological degeneration and tibiofemoral osteophytes [72].

However, as an operator-dependent modality, ultrasound cannot visualize subchondral bone changes and can visualize only parts of the articular chondral surface such as cartilage primarily within the patellofemoral joint [22]. Consequently, up to now, in the field of knee OA analysis, ultrasound has just dabbled in preprocessing and segmentation. Md Belayet Hossain et al. modified histogram equalization by finding out the separating point for segmenting histogram for which brightness and detail preservation would be achieved while enhancing the contrast at the same time [71]. [70] enhanced the low-intensity bone surfaces and cartilage interface by constructing a local phase-based enhancement metric, followed by Random-Walker (RW) algorithm, a graph-based segmentation scheme.

| OA features | Radiograph | MRI | CT | Ultrasound |
|---|---|---|---|---|
| Cartilage | + | ++++ | +++ | ++ |
| JSN | ++ | +++ | +++ | + |
| Subchondral cysts and sclerosis | ++ | +++ | ++++ | - |
| Bone marrow lesions | - | ++++ | ++ | - |
| Osteophytes and erosions | ++ | +++ | ++++ | ++ |
| Inflammation | - | ++++ | + | +++ |
| Soft tissues (menisci, tendons) | - | ++++ | ++ | +++ |

**Table 2.7:** Modality detection performance comparison in OA diagnosis and follow-up

Even if ultrasound has been increasingly deployed for the assessment of hand OA, they still adopt KL and OARSI scores instead of creating a more appropriate semi-quantitative metric [22]. Ultrasound hasn't yet accomplished evaluation metrics establishment and workflow exploration, which also substantiates the compatibility of radiographic semi-quantitative criteria.

## 2.5 Modality comparison

Demonstrated in Table 2.7 [16], Table 2.8 [16] and Table 2.9, as the two relatively mature modalities, no matter from which aspects (detected OA features, clinical utility and workflow emphasis), radiography and MRI are complementary. Following above two modalities, CT, ultrasound and other rising modalities, such as vibroarthrography [73], spectroscopic images [74] and so on, have set out to inaugurate automatic knee OA prediction systems from both medical experiments like assessment standard formulation and informatics researches such as eligible quantification framework selection.

| Clinical utility | Radiograph | MRI | CT | Ultrasound |
|---|---|---|---|---|
| Early diagnosis | + | +++ | +++ | +++ |
| Feasibility | ++++ | +++ | ++ | +++ |
| Cost | ++++ | ++ | ++ | +++ |
| Radiation dose | ++ | ++++ | ++ | ++++ |
| Data complexity | + | ++++ | +++ | ++ |
| Quantitative metrics | ++++ | +++ | - | - |

**Table 2.8:** Modality clinical utility comparison in OA diagnosis and follow-up

| Stage | Radiograph | MRI | CT | Ultrasound |
|---|---|---|---|---|
| Preprocessing | + | ++++ | +++ | ++ |
| ROI extraction | ++ | ++++ | ++ | ++ |
| Classification | ++++ | ++ | +++ | - |

**Table 2.9:** Modality research focus comparison in automatic knee OA diagnosis

The more "+" represents the better performance, whereas "-" stands for the lack of related ability/research, which is the same with Table 2.7 and Table 2.8.

Confronted with the maturity of radiography and MRI, the potential of those different modalities remains elusive, in particular for long-term data collection.

# Chapter 3

# Multimodality Based Automatic Knee OA Quantification

As described earlier, each modality has its respective virtues and bottlenecks, which arouses the exploration for the combination of modalities. After all, reciprocal modalities can mutually reinforce detected features by drawing each other merits. In addition, plenty of algorithms have confirmed the significant enhancement of robustness and accuracy by classifier ensemble [8, 35, 75, 76]. Compared with reconstructing voxel data from 2D slices for 3D classification, model ensemble is prevented from "curse of dimensionality", which diminishes accuracies by demanding a great deal of training data [77].

The scaling the amount of data exponentially is unrealistic for medical images. As a consequence, in line with the detection frameworks of individual modalities, the desirable multimodality based automatic knee OA quantification approach is developed as depicted in fig. 3.1. Meanwhile, the complementary, accessibility and maturity analyzed in chapter 2 settle radiography and MRI as experimental candidates.

## 3.1 Preprocessing

The substantial divergence of their data complexity results in the operation distinction of radiograph and MRI preprocessing. In general, rescaling and horizontal flipping constitute the entire preprocessing of radiographs, whereas MRIs yet call for contrast enhancement, intensity regulation, noise elimination and different measurement integration. Merging their ultimate demands, the preprocessing steps are portrayed in fig. 3.2. Not only owing to the requirement of input dimension alignment from the further deep learning model training, as the most primitive operation of image preprocessing, rescaling to smaller size is also driven by the limited memory of training devices, in particular

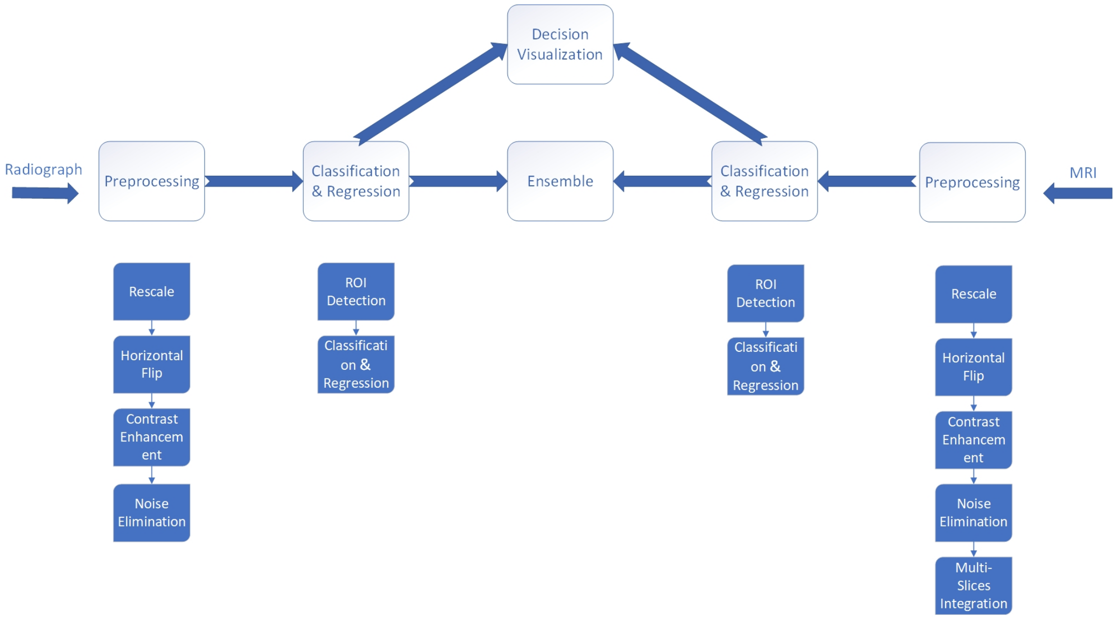


**Figure 3.1:** Workflow of the proposed approach

for those high-resolution radiographs.

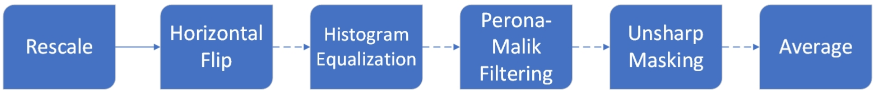


**Figure 3.2:** The general preprocessing pipeline

Differing from the data augmentation aim of previous papers [8, 28], for the sake of classifier error reduction, all the knee images are horizontally flipped into the same direction. Apart from the foregoing two basic operations, contrast enhancement and noise elimination are also corroborated to be universally propitious to both modalities. As for intensity regulation, it is roughly designed for MRI segmentation based on edge detection, which strays from the state-of-art technical tendency. Moreover, up to now, contrast enhancement is carried out as intensity normalization. Accordingly, intensity adaption is incorporated into contrast enhancement step. Compared with global research data collections, in clinical practice, MRI machines won't adjust settings per patient, which determines the impossible frequent occurrence of the multiformity of parameters. Certainly, the diversity of slices from MRI sequences still comes into preprocessing focus, which deserves special treatments.

### 3.1.1 Contrast enhancement

As one of the most determinant factor for image quality, integral contrast is defined as the difference in the pixel intensity value of a particular pixel to its neighboring pixels [78]. The more contrast gives better clarity of an image in terms of local details. Aiming to highlight the interpretability and perception of information contained in images, especially in details of dark regions, contrast enhancement acquires clear images through brightness intensity value redistribution by means of stretching interval between dark and brightness area without significant distortions [79].

In general, the acknowledged approaches for contrast enhancement are subdivided into two types in accordance with the image stretching span namely global and local methods [78]. The global one, including histogram equalization as well as its modifications, non-linear stretching (logarithmic, exponential and power functions, etc.) and adaptive linear stretching, mainly focuses on overall viewing purpose, which leads to a possible disappearance of small-size objects on images, especially when images contain regions expressly darker or brighter than other parts [78, 80, 81, 82]. Obviously, a better way to address such problem is to enhance the dark regions by keeping the bright regions untouched, where local contrast enhancement is inspired [78]. However, combining a small area or neighbourhood of pixels to generate an enhanced pixel is not an acceptable image processing in real time [80, 83]. Thus, based on saliency principle that human vision is sensitive to high frequency contents, instead of regions, the academic gradually gives prominence to edge emphasis, which is also named as image sharpening. Contrary to global contrast enhancement, image sharpening would generate magnified noises, which results in the operation order exchange that edge enhancement should be operated after noise elimination, as shown in fig. 3.3.

There is no universal methods so far and the specific methodology depends on the context of related tasks and image content. Roughly, local contrast enhancement is too computationally intensive for medical images, which are relatively large and high-resolution. Hence, this thesis only draws an all-round comparison between global contrast enhancement and image sharpening. As depict in chapter 2, compared to radiography, MRI displays more detailed OA features, such as bone marrow lesions, inflammation, soft tissues and so on, which, to a great extent, would vanish due to global contrast enhancement. In this case, as exhibited in fig. 3.4, global operations are not employed to MRIs in our evaluation.

#### Global contrast enhancement

Having witnessed to its triumph in the domain of medical images [11, 14, 52, 53, 54, 82, 83], rather than non-linear and adaptive linear stretching, typical histogram equalization is assigned as the representative of global contrast enhancement in the following evaluation. An image histogram is the graphical representation of the relative frequencies of the

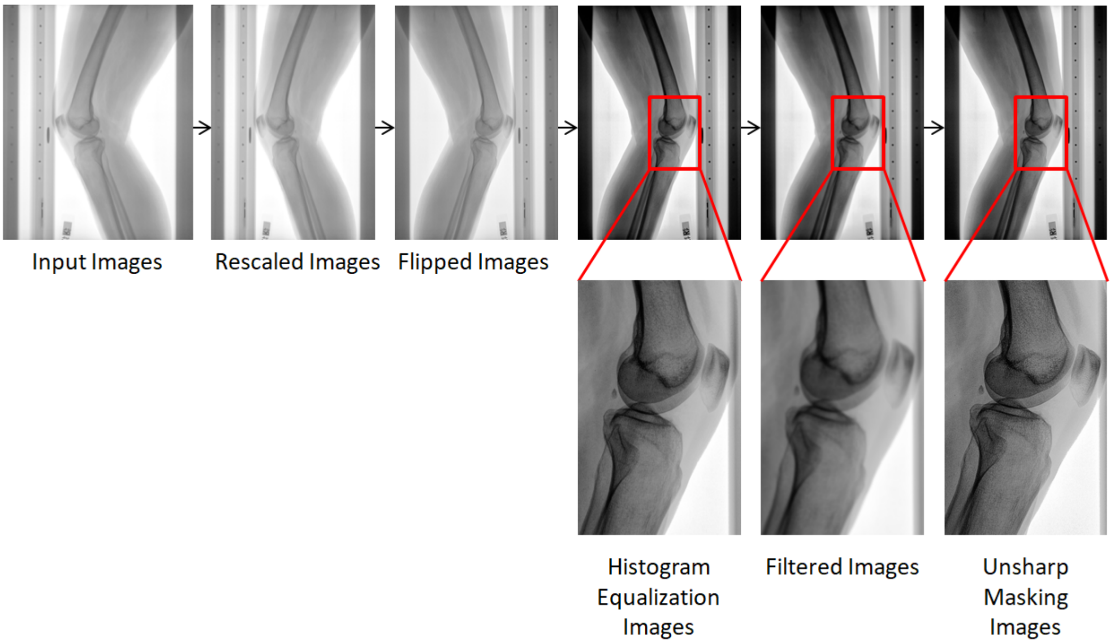


**Figure 3.3:** The preprocessing pipeline of radiographs

different gray levels in that image, which provides a total description of the appearance of an image [82]. The basic concept of histogram equalization lies on mapping gray levels from their original intensity probability distribution to a uniform distribution, which flattens and stretches the entire dynamics range of the image histogram resulting in overall contrast modification [82, 84].

In other words, via combining gray levels with less frequencies into one and stretching high frequent intensities over high range of gray levels, histogram equalization achieves close to equally distributed intensities [82, 84]. After all, information entropy will be at peak, when data have uniform distribution property [84]. To be specific, the probability density function $p(X_k)$ of a given image $\boldsymbol{X}$ is defined as eq. (3.1) [84]

$$p(X_k) = \frac{n_k}{N} \tag{3.1}$$

where $k$ is the gray level ID of input image $\boldsymbol{X}$ varying from 0 to $L$ and $n_k$ represents the frequency of gray level $X_k$ appearing in $\boldsymbol{X}$. As for $N$, it is the total number of samples from the input image $\boldsymbol{X}$. Therefore, a plot of $n_k$ vs. $X_k$ is specified as the histogram of $\boldsymbol{X}$, while the equalization transform function $f(X_k)$ is tightly related to the cumulative density function $c(X_k)$:

$$f(X_k) = X_0 + (X_L - X_0)c(X_k) \tag{3.2}$$

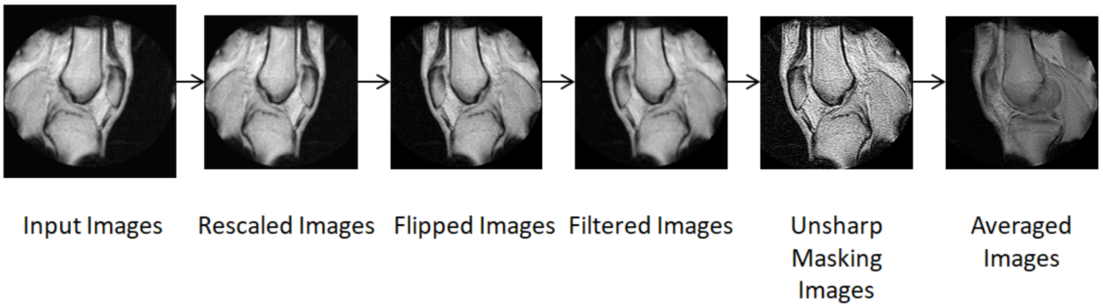


**Figure 3.4:** The preprocessing pipeline of MRIs

$$c(X_k) = \sum_{j=0}^{k} p(X_j) \tag{3.3}$$

Synthesizing above formula derivation, the output of typical histogram equalization $\boldsymbol{Y} = Y(i, j)$ should be expressed in the following:

$$\boldsymbol{Y} = f(\boldsymbol{X}) = \{f(X(i, j)) | \forall X(i, j) \in \boldsymbol{X}\} \tag{3.4}$$

### Edge enhancement

Current available techniques of edge enhancement can be broadly classified into two categories: spatial domain and frequency domain [85]. The former directly operates on pixels, which favours real time implementations, thanks to the conceptual briefness. However, this leads to the lack of robustness, in particular from imperceptibility perspective [85]. In order to figure out this dilemma, frequency-based image sharpening is blazed, which manipulates image transform coefficients after Fourier transform. Evidently, it doesn't carry forwards the low complexity of computation and the ease of viewing. In the meanwhile, the full images cannot be simultaneously tickled [85]. Thus, this thesis only culls one typical methodology from spatial domain.

Owing to the spread of image processing software and libraries, such as Adobe Photoshop, Python Imaging Library (PIL) and so on, there is no doubt that unsharp masking is the most mature technique for image sharpening. In general, the blurry unsharp masking is derived from a weighted highpass-filtered version of the original image, which would be added back to the signal itself, as shown in fig. 3.5 and eq. (3.5) [86].

$$Y = X + \lambda * g(X) \tag{3.5}$$

Referring to the implementation of PIL, ImageFilter.EDGE_ENHANCE and Im-

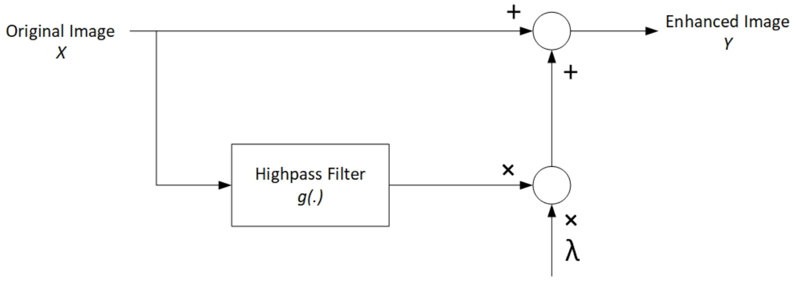
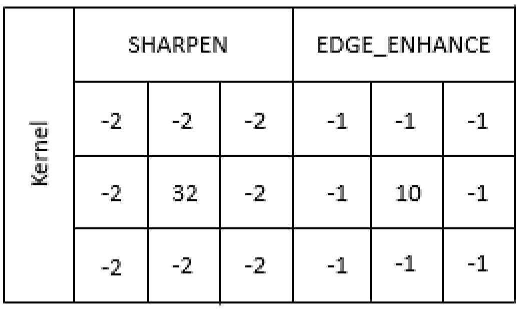


**Figure 3.5:** The working principle of unsharp masking

ageFilter.SHARPEN are adopted as the kernel ($g(.)$) for further evaluation, whose convolution matrices are demonstrated in fig. 3.6. After all, by contrast, another kernel ImageFilter.EDGE_ENHANCE_MORE magnifies too many noises in practice.



| | SHARPEN | | | EDGE_ENHANCE | | |
|---|---|---|---|---|---|---|
| Kernel | -2 | -2 | -2 | -1 | -1 | -1 |
| | -2 | 32 | -2 | -1 | 10 | -1 |
| | -2 | -2 | -2 | -1 | -1 | -1 |

**Figure 3.6:** The kernel of EDGE_ENHANCE and SHARPEN

Considering the compatibility between histogram equalization and image sharpening, apart from single contrast enhancement approach, the combinations of histogram equalization and different sharpening kernels are also inclusive in the preprocessing comparison scheme.

### 3.1.2 Noise elimination

Noise is a random variation of image intensity, which is visible as grains in images. In the light of the characteristics of diverse noises, heretofore, medical images undergo sundry smoothing filters for common noise removal, chiefly including median filter, Gaussian filter and Perona-Malik filter. Median filter is born for impulse noise (named also as Salt-and-Pepper noise) which is provoked by the sharp and sudden disturbance in the image signal or transmission error [87]. Apparently, distinct from normal pictures, above two causes rarely occur in medical images. Therefore, our OA quantification strategy doesn't take median filter into consideration.

As expounded in chapter 2, in comparison to Gaussian filter, profuse empirical evaluations have substantiated the superiority of Perona-Malik filter which preserves edges and detailed structures along with noise reduction, as long as the fitting diffusion coefficient $c(.)$ and gradient threshold $K$ are singled out [14, 88]. Hence, our noise elimination exploration concentrates on the parameter selection of Perona-Malik filter. As a non-linear anisotropic diffusion model, Perona-Malik filter smoothens noisy images $\theta(x, y)$ by means of the partial differential equation [88]:

$$\frac{\partial u}{\partial t} = div(c(|\nabla u(x,y,t)|)\nabla u(x,y,t)) \tag{3.6}$$

where $u(x, y, t)$ serves as the obtained image after $t$ iteration diffusion. Videlicet, $u(x, y, 0)$ is the original noisy image $\theta(x, y)$. Moreover, $div$ and $\nabla$ correspondingly indicate the divergence operator and the gradient operator with respect to the spatial variables $x$ and $y$. As for the diffusion coefficient $c(.)$, the initial authors Perona and Malik nominated below two functions [89]:

$$c_1(|\nabla I|) = exp\left(-\left(\frac{|\nabla I|}{K}\right)^2\right) \tag{3.7}$$

$$c_2(|\nabla I|) = \frac{1}{1+\left(\frac{|\nabla I|}{K}\right)^2} \tag{3.8}$$

Together with Tukey's biweight function (manifested as shown in eq. (3.9)) [90], they compose all the eminent diffusion coefficients. Definitely, their outcomes are inversely proportional to the magnitude of the local image gradient. Accordingly, within inner regions, the gradient magnitude is weak, thereby the diffusion coefficient is almost 1, which acts as typical heat equation to smoothen the relative regions. In contrast, the strong gradient of boundaries engender the diffusion stop due to the nearly zero diffusion coefficient. Regarding whether the local gradient magnitude is strong enough for edge preservation, apart from gradient threshold $K$, it depends on which diffusion coefficient function $c(.)$ is picked as follows:

$$c_3(|\nabla I|) = \begin{cases} \frac{1}{2}\left[1-\left(\frac{|\nabla I|}{K\sqrt{2}}\right)^2\right]^2, & |\nabla I| \le K\sqrt{2} \\ 0, & |\nabla I| > K\sqrt{2} \end{cases} \tag{3.9}$$

The Tukey's biweight function $c_3$ is once recognized as the best option, since its boundary between noises and edges is lowest [88]. However, under the circumstances, noises are also likely to be left. Conversely, $c_2$ has the highest threshold to distinguish noises and edges so that conspicuously sharp edges and fine details would be diffused. Thus,

$c_1$, which thoroughly conserves edges and highlights in conjunction with general noise reduction, is adopted in this thesis. Concerning the remained two parameters, gradient threshold $K$ and number of iterations $t$, our further OA diagnosis probe would lay more emphasis on their selection from practical aspect.

### 3.1.3 Multi-slices integration

Different with the multiformity of pulse sequences and parameters in research MRI collections, the diversity of slices becomes the heart of how to wield MRI sequences in clinics. Consequently, the traditional Sobel filter and affine transformation cannot reveal their talent, since the shapes of knees are even erratic.

Referring to the application of MRIs in other body parts [91, 92], our scheme makes use of average filter which acquires the numerical mean of corresponding pixels for multi-slices so that MRI slices are integrated whilst further erases noises and artefacts at the single MRI slice. Absolutely, in order to avoid strong bias, average operation takes place after discarding certain images in the beginning and the end of series. For the sake of validating the promotion of average filter, this thesis still picks up MRI slices in the middle of the entire sequences as control trail.

## 3.2 Classification and regression analysis

According to the proof in [34], regression analysis was certified to promote the accuracy so that our blueprint introduces homologous regression algorithms based on multiclass classification. Hereto, summarizing previously stated architectures in the literature, there are three types of quantification workflows:

- Type A (ML-based approach): ROI detection + feature extraction + classification/regression
- Type B (CNN-based approach): ROI detection + classification/regression
- Type C (end-to-end framework): Classification/Regression.

However, compared with Type A, [34] has manifested the preeminence of Type B, by virtue of CNN's predominantly automatic feature extraction. Hence, this research is centred on Type B (CNN-based approach) and Type C (end-to-end framework) architectures.

### 3.2.1 CNN-based approaches

Conducive to weakening the negative influence of image artifacts on classification, traditional CNN prediction workflow regularly starts with ROI extraction to glean the knee-joint rectangles where own the most decisive features from the full medical images. Within those pivotal regions, a couple of CNN variants applied to general images are opted to fulfill grading through classifying hierarchies of features without resorting to feature engineering schemes [93]. Distinctly, labelling massive images manually where ROIs are is so drained that drawing on recent successes of deep learning in semantic segmentation, automatic ROI detection phase is also consummated by CNN branches.

#### ROI extraction

So-called ROI extraction conventionally trains neural networks by ground truths as labels which specify ROIs with masked binary images or intrinsically 2D matrices in the same size of input images. Broadly, if related pixels are within ROIs, ground truths mark as 1, whereas the rest tags as 0. Having trained relative models, the bounding boxes of ROIs can be deduced in accordance with the contour coordinates simply acquired from model predictions as fig. 3.7. In consequence, this subsection underlines the most crucial step to eventually attains fully automatic ROI extraction, segmentation model training.

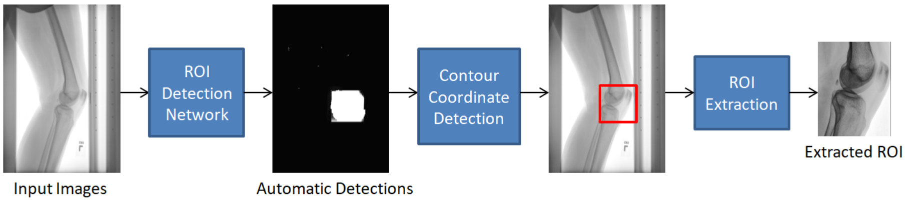


**Figure 3.7:** Workflow of the ROI extraction process

As interpreted in chapter 2, the breakthrough from [34] to [28] has shown the power of FCN in the ROI detection domain so that our evaluation would implement FCN declared in [28] as baseline. The basic concept behind FCN roots in pixel-wise prediction, whose key to success is to leverage large-scale image classification as supervised pre-training and fine-tune fully convolutional layers via transfer learning [93]. As an end-to-end semantic segmentation technique without further machinery, FCN adapts standard deep CNN to learn per-pixel labels from ground truths of entire images by removing fully connected layers as well as importing deconvolution layers [93, 94]. Purely with a train of non-linear filters, FCNs naturally map coarse outputs to the dense pixel space [93, 94]. To be specific, the final layer generates tensors in the consistent spatial dimensions of inputs, except the number of channels will be equal to the number of classes (customarily 2 classes: ROI and non-ROI) whose likelihoods would be calculated by softmax probability function at the same time.

Undoubtedly, striding and pooling diminish image dimensions, where deconvolution or essentially transposed convolution layers treated as inverse operation realize their potentials to match the widths and heights of the original input images. The transposed convolution is implemented as the backward pass of related convolutional operator with regards to weights, which can be effectuated via two procedures: zero padding and unit strides [95]. The former approach executes filter dilating by padding zeros between adjacent filter elements and cross-correlates them with the input [94, 95]. The latter achieves upsampling with factor $f$ by convoluting with a fractional input stride $1/f$, where the alternative name, fractional stride convolution, is designated [94]. Aside from transposed convolution, interpolation is also supposed to be the most efficient technique for upsampling. Undoubtedly, rather than fixed interpolation, deconvolution filter can be trained, even as a non-liner upsampling. Thus, transposed convolution is applied to our design.

Similar with other semantic segmentation architecture, it faces an inherent tension between semantics and location: global features resolve semantics while local information unveils positions. Deep feature hierarchies jointly encode location and semantics in a local-to-global pyramid so that FCN defines combination layer by element-wise addition for fusing deep, coarse, semantic features and shallow, fine, appearance features [94]. Conforming to the initial FCN network, there are three modes for target mixture: FCN-32s, FCN-16s and FCN-8s, as presented in fig. 3.8 [96].

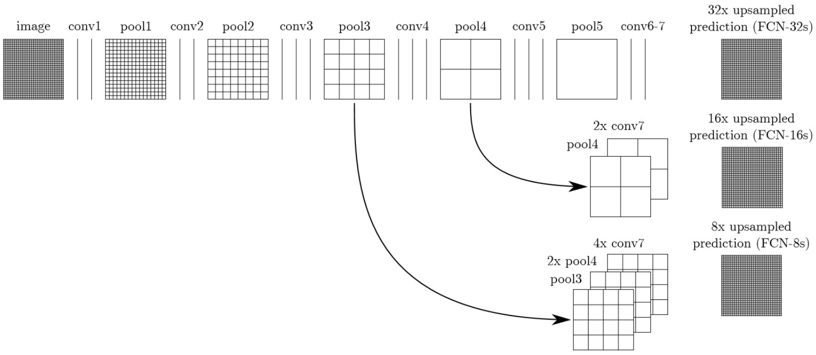


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled prediction (FCN-32s)
2x conv7
pool4
16x upsampled prediction (FCN-16s)
4x conv7
2x pool4
pool3
8x upsampled prediction (FCN-8s)

**Figure 3.8:** The FCN architecture

Without any fusion, FCN-32s directly upsamples the output of last convolutional layer at stride 32, which loses predominantly spatial information [94]. In line with above theoretical analysis, FCN-16s and FCN-8s are developed on the basis of what fig. 3.8 displays. FCN-16s adds the output of penultimate pooling layer and 2 × upsampled prediction from last convolutional operation, whose combination then performs 16 × upsampling. Regarding FCN-8s, the sum obtained from 2 × upsampled production of last convolutional layer (with a stride 2 transposed convolution) and the output of penultimate max-pooling links with the antepenultimate pooling production. Certainly, in order to accomplish element-wise addition, the previous sum has to enlarge as twice size. On the top of this combined feature map, a transposed convolution layer with stride 8 is carried out for the final segmentation map.

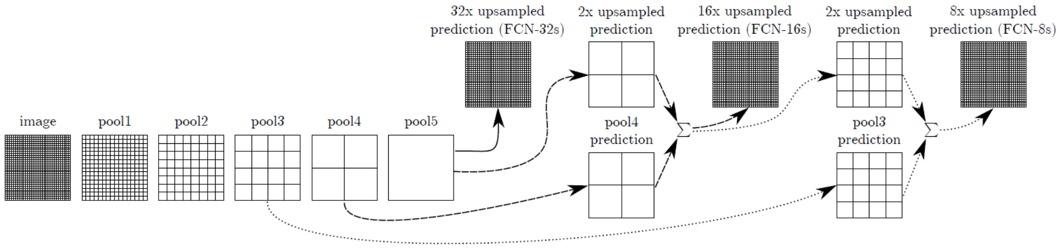


**Figure 3.9:** The FCN fusion principle

Visibly, as illustrated in fig. 3.9 [94], our baseline FCN noted in [28] (fig. 3.10) belongs to FCN-32s. However, [94] has substantiated FCN-8s delivered the best performance from both theoretical and empirical perspective. Therefore, we establish a lightweight FCN-8s network with the same settings from scratch to upgrade ROI detection performance. The network consists of 4 convolution blocks followed by a max-pooling layer per block and 3 upsampling stages. The kernels of convolution and max-pooling are uniformed as [3×3] and [2×2] severally, whereas the numbers of filters for convolutional operators at each stage are rising in the way: 32, 32, 64 and 96. Naturally, due to vanishing gradient prevention, batch normalization and Rectified Linear Unit (ReLU) activation function also come with each convolution layer. In relation to upsampling stages, the first two have the same [2×2] stride transposed convolution attaching an add layer, while the last stage merely deconvolutes at [4×4] stride with Sigmoid activation function, whose concatenate details are indicated in fig. 3.11.

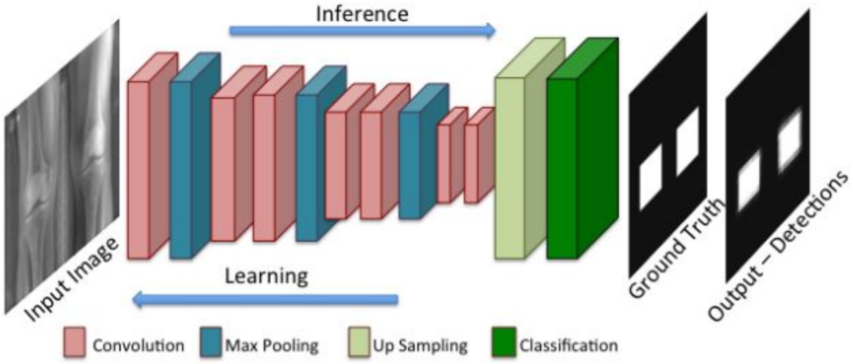


**Figure 3.10:** The baseline FCN architecture

Supplementing a usual contracting network by successive layers, where pooling opera-

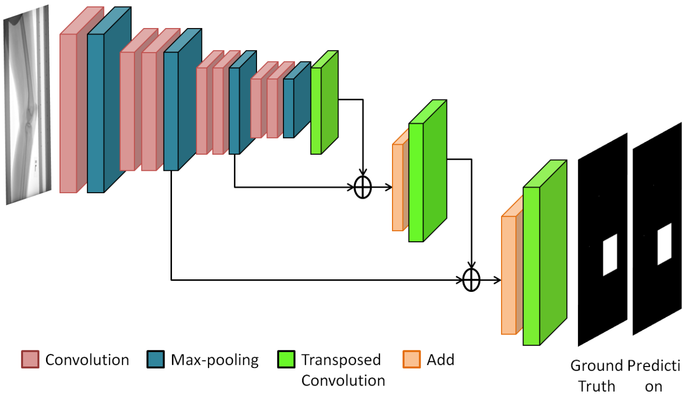


**Figure 3.11:** Revised FCN-8s architecture

tors are replaced by upsampling operators, FCN reinforces the output resolution [97]. In order to assemble a more precise segmentation, high resolution features from the contracting path of FCN are combined with the upsampled outcomes [97]. Imaginably, feature channels in the expansive path are augmented, which allow the network to propagate the full stack of context available in local-to-global pyramid hierarchies to higher layers. To a great extent, it outperforms the prior best mechanism, for which a more elegant network, U-Net, is come up with. Namely, U-Net yields a u-shaped architecture whose expansive path is symmetric to the contracting path as shown in fig. 3.12 [97].

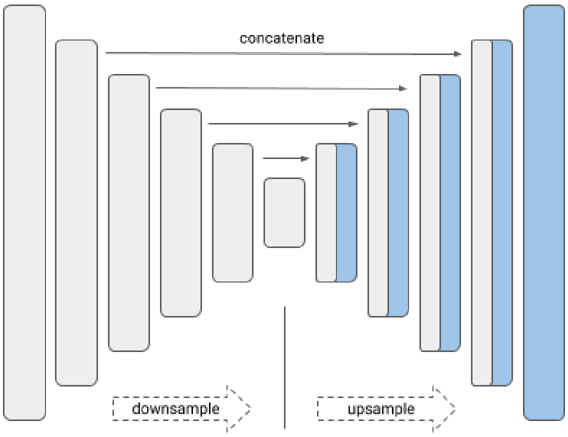


**Figure 3.12:** The U-Net architecture

As the winner in biomedical image segmentation category by a large margin, beyond the symmetric structure, U-Net has a brilliant trick up its sleeve: concatenation operators instead of element-wise addition. These skip connections intend to provide local information to the global context while upsampling so that the decoder at each stage owns all the relevant features that are lost when pooled in the encoder. Another stroke of genius from U-Net is no padding in convolutional layers [97]. Whereupon, only valid feature maps are left after convolutions, which brings in the seamless segmentation of arbitrarily large images by an overlap-tile strategy that to predict the pixels in the border regions of images, the missing context is extrapolated by mirroring the input images [97]. This tiling strategy is the key why U-Net perfect matches with medical image segmentation, which is always confined by GPU memory, due to the huge input size [97].

The only pitfall needs to be noted that in this case, cropping feature maps from the contracting path is indispensable, arising from the loss of border pixels in every convolution [97]. The prevalence of U-Net is chiefly beneficial to the more flexible contracting path, which could follow any typical convolutional network (also termed as backbone), or even self-tuned architecture. Consequently, the inner drive to succeed of U-Net is how to sort out an ideal backbone. In view of vanishing gradient avoidance and faster training convergence verified in [98], ResNet is extended to our U-Net.

While U-Net is proud of its long skip connections between contracting and expansive path, ResNet is glorious for its shortcut connections among convolutional layers. With the depths of neural networks growing, the degradation of training accuracy issue has been exposed. ResNet provides two mappings to settle down this challenge: identity mapping ($x$) and residual mapping ($F(x)$) (shown in fig. 3.13 [99]), where the architecture is named after [99]. The outcome of each residual block $y$ then is $F(x) + x$. To the extreme, if the network training is optimal, solvers simply drive the weights of the multiple non-linear layers toward zero to approach identity mappings so that the network would maintain in the best status, even if architectures are deeper and deeper [99].

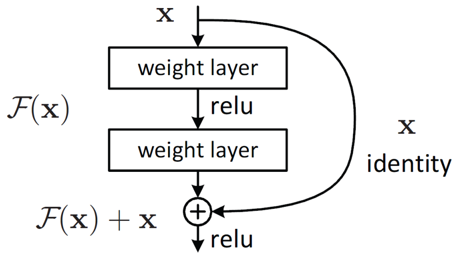


**Figure 3.13:** The residual block design

Although shortcuts don't step up extra parameters, CNN deepening still leads to the parameter explosion. Accordingly, apart from common building blocks as fig. 3.14 left, [99] also designed bottleneck (fig. 3.14 right) by means of channel reduction. The former generally constructs ResNets with 34 or less layers, while the latter is invented for ResNets with over 50 layers.

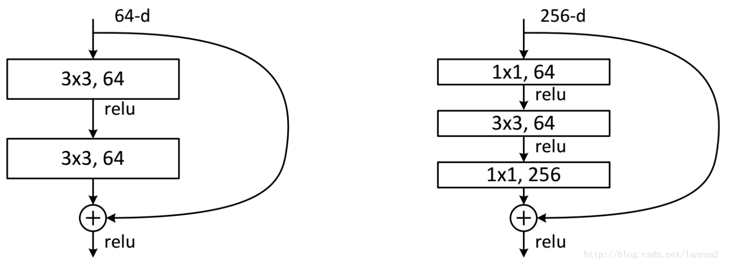


**Figure 3.14:** Two types of residual blocks

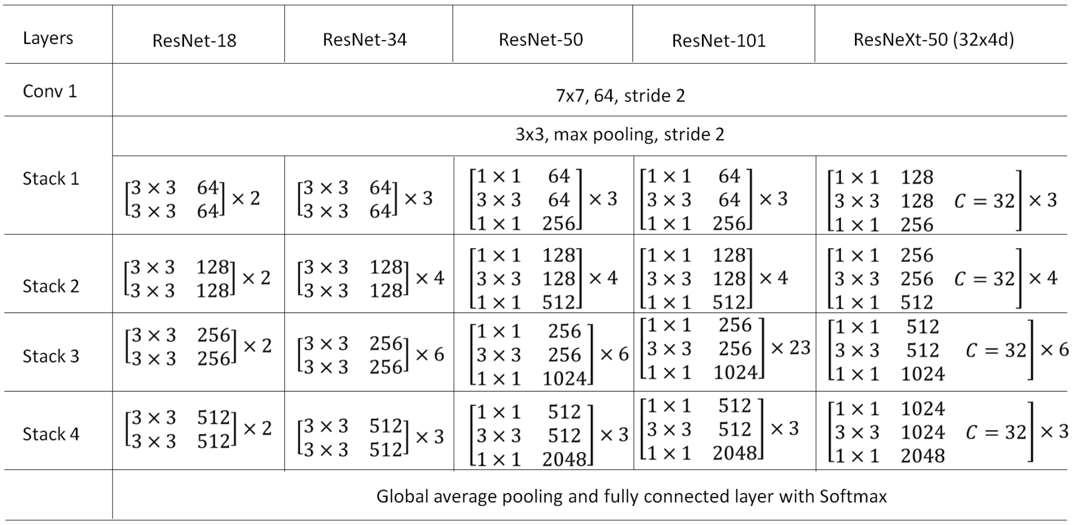
A series of convolution operators without pooling in between is recognized as a stack. The first convolutional layer of each stack (except the first stack) in ResNets downsamples by at stride 2, which inevitably provokes the channel difference between identity mapping ($x$) and residual mapping ($F(x)$). Under this situation, the output of each block $y = F(x) + Wx$, where $W$ is also a convolution operator for $x$'s channel adjustment [99]. Recently, ResNet almost evolves into the first alternative of classification tasks, for which there are plentiful variants thanks to the ample layer combinations. Served as backbone of U-Net, excessive layers would impose more computing burden. Hence, only ResNet-18, ResNet-34, ResNet-50 and ResNet-101 (detailed architectures shown in fig. 3.15 [99, 100]) are taken into consideration.

Exactly, architectures with superabundant hyper-parameters, like filter sizes, channels and so on, also suffer from complicated computing and memory exhaustion so that ResNeXt-50 (network settings displayed in fig. 3.15 [99, 100]) is involved into our comparison as well. ResNeXt is the modified ResNet with homogeneous and multi-branch architecture that has a few hyper-parameters to assign by repeating a building block that aggregates a set of transformations with the same topology as fig. 3.16 [100]. fig. 3.17 [100] clearly exhibits the equivalence relation of ResNet and ResNeXt blocks, which confirms that even under the restricted condition of maintaining network architecture, ResNeXt can handle with higher hyper-parameter settings.

Nevertheless, no matter which backbone is employed, since all the networks are comprised of a single convolutional layer and 4 stacks, their concatenations are fixed to lay out before each stack, likewise what is manifested in fig. 3.18.

### Classification and regression in ROIs

OA grading for ROIs overtly pertains to CNN, in particular special networks, such as VGG-m-128, Siamese CNN, ResNet-34 and DenseNets. Referring to their reputation in the sphere of image classification, further quantification evaluation would build up VGG, ResNet and DenseNet series. Considering the strategies and layouts of investigated ResNet variants: ResNet-18, ResNet-34, ResNet-50 and ResNet-101 have been set



| Layers | ResNet-18 | ResNet-34 | ResNet-50 | ResNet-101 | ResNeXt-50 (32x4d) |
|---|---|---|---|---|---|
| Conv 1 | 7x7, 64, stride 2 | | | | |
| Stack 1 | 3x3, max pooling, stride 2 | | | | |
| | $\begin{bmatrix} 3\times3 & 64 \\ 3\times3 & 64 \end{bmatrix}\times 2$ | $\begin{bmatrix} 3\times3 & 64 \\ 3\times3 & 64 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 64 \\ 3\times3 & 64 \\ 1\times1 & 256 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 64 \\ 3\times3 & 64 \\ 1\times1 & 256 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 128 & \\ 3\times3 & 128 & C=32 \\ 1\times1 & 256 & \end{bmatrix}\times 3$ |
| Stack 2 | $\begin{bmatrix} 3\times3 & 128 \\ 3\times3 & 128 \end{bmatrix}\times 2$ | $\begin{bmatrix} 3\times3 & 128 \\ 3\times3 & 128 \end{bmatrix}\times 4$ | $\begin{bmatrix} 1\times1 & 128 \\ 3\times3 & 128 \\ 1\times1 & 512 \end{bmatrix}\times 4$ | $\begin{bmatrix} 1\times1 & 128 \\ 3\times3 & 128 \\ 1\times1 & 512 \end{bmatrix}\times 4$ | $\begin{bmatrix} 1\times1 & 256 & \\ 3\times3 & 256 & C=32 \\ 1\times1 & 512 & \end{bmatrix}\times 4$ |
| Stack 3 | $\begin{bmatrix} 3\times3 & 256 \\ 3\times3 & 256 \end{bmatrix}\times 2$ | $\begin{bmatrix} 3\times3 & 256 \\ 3\times3 & 256 \end{bmatrix}\times 6$ | $\begin{bmatrix} 1\times1 & 256 \\ 3\times3 & 256 \\ 1\times1 & 1024 \end{bmatrix}\times 6$ | $\begin{bmatrix} 1\times1 & 256 \\ 3\times3 & 256 \\ 1\times1 & 1024 \end{bmatrix}\times 23$ | $\begin{bmatrix} 1\times1 & 512 & \\ 3\times3 & 512 & C=32 \\ 1\times1 & 1024 & \end{bmatrix}\times 6$ |
| Stack 4 | $\begin{bmatrix} 3\times3 & 512 \\ 3\times3 & 512 \end{bmatrix}\times 2$ | $\begin{bmatrix} 3\times3 & 512 \\ 3\times3 & 512 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 512 \\ 3\times3 & 512 \\ 1\times1 & 2048 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 512 \\ 3\times3 & 512 \\ 1\times1 & 2048 \end{bmatrix}\times 3$ | $\begin{bmatrix} 1\times1 & 1024 & \\ 3\times3 & 1024 & C=32 \\ 1\times1 & 2048 & \end{bmatrix}\times 3$ |
| | Global average pooling and fully connected layer with Softmax | | | | |

**Figure 3.15:** Different ResNet-based architectures

Parameters of convolutional layers are denoted as [receptive field size & the number of channels] × repetition times, where $C$ represents the path number of repeated residual blocks and $4d$ points out the channels of each aggregative building block. Downsampling is performed at the first layer of Stack 2, Stack 3 and Stack 4 with a stride of 2.

forth in section 3.2.1, here we spotlight VGG and DenseNet.

Essentially, ResNet is inherited from VGG concept, a lightweight stack-based CNN. Compared with structures in the early stage, VGG ushers CNN deepening, thanks to its very small (3×3) convolution filters [76]. In terms of regularisation definition, a 7×7 convolutional filter can decompose into a stack of three 3×3 filters with non-linearity injected in between [76]. Plainly, in the meanwhile of layer extension, incorporating three non-linear ReLU activation functions instead of a single one makes the decision equation more discriminative and parameter shrinkage [76].

Karen Simonyan and Andrew Zisserman in [76] proposed 6 VGG networks equipped with 3×3 filters and 2×2 max-pooling as shown in fig. 3.19 [76], in which VGG-16 and VGG-19 (D and E) stand out from pragmatic image classification tasks. Therefore, this thesis lays more attention on these two architectures for our OA scoring.

Certainly, different with inchoate CNNs, VGG consumes much more computing resources caused by the three fully connected layers which do has not so much impact on performance gain. Our first modification on VGG-16 and VGG-19 is hidden node reduction for the first two fully connected layers, which would be explored via detail medical images. After all, the design of 4096 hidden nodes are caused by 1000 classes in ILSCRC dataset. Referring to dense evaluation concept presented in [76], the other VGG revision converts the last three fully connected layers to 2D convolution operation for parameter

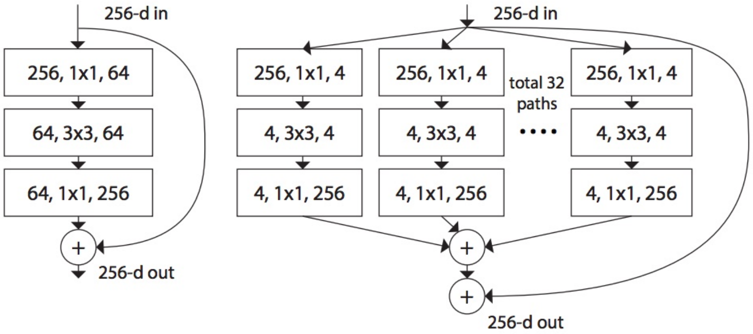


**Figure 3.16:** Block comparison in ResNet and ResNeXt

Left: ResNet block shown as number of in-channels, filter size and number of out-channels; Right: ResNeXt block with roughly the same complexity

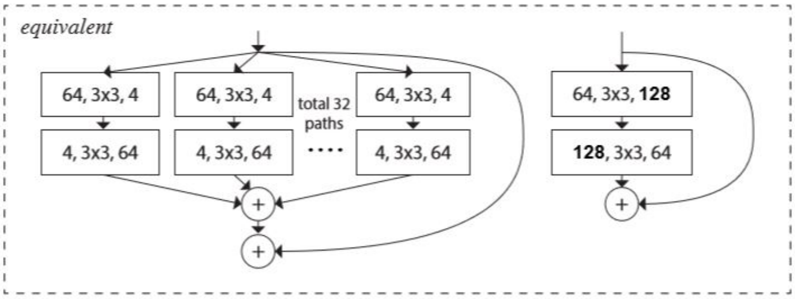


**Figure 3.17:** Block relationship between ResNet and ResNeXt

Left: ResNeXt block; Right: Equivalent ResNet block

decline. For the sake of classification, the last two layers are fixed to adopt 1×1 kernel but only the final one is equipped with Softmax activation function, likewise fig. 3.20. As for the remaining settings, they leave the potential space for the following empirical exploration.

As ResNet embraces the originality of VGG, DenseNet distills the shortcut insight of ResNet to surpass the 150-layer barrier, which direct connects from any layer to all subsequent layers in a feed-forward fashion [101]. Apparently, it carries forward the vanishing-gradient alleviation by a simple connectivity pattern to ensure maximum information flow between layers in the network [101]. After all, each layer has direct access to the gradients from the loss function and the original input signal, contributing to an implicit deep supervision [101]. ResNet merges feature-maps through summation, while DenseNet concatenates additional inputs from all preceding layers in channel dimension, which not only strengthens feature propagation to moderate information loss, but also encourages feature reuse to substantially cut down the number of parameters from width [101]. Since there is no need to relearn redundant feature-maps, DenseNet can be very narrow, which depends on the growth rate $k$, defining how many feature-maps from each layer needed to be collected for the final decision [101]. On account of downsampling, concatenation cannot be executed straightforwardly, for which transition layers, amount to a batch normalization layer, a $1\times1$ convolutional layer and a $2\times2$ average pooling layer, are injected into dense blocks as shown in fig. 3.21 [101].

In light of above compelling advantages, even regularization effect observed from experiments in [101], DenseNet can freely breakthrough the 100-layer barrier. Hence, current mainstream DenseNet topologies are drafted from 121 layers to 201 layers (shown in fig. 3.22), which are all involved into our OA prediction scheme, thanks to their strong possibilities of lifting up classification accuracy by depth.

In conclusion, by reference of theoretical structure analysis and previous performance rankings, our traditional CNN-based approach is selected from the following combinations of ROI detection and classification/regression (fig. 3.23).

### 3.2.2 End-to-end neural networks

The soaring popularity of end-to-end learning triggers our interest for exploration in the medical image classification realm. After all, it has already yielded extraordinary talents on object detection. In order to search for classes of all the targets in a picture, traditional CNN divides images into separated regions. Having been classified each region, the original image merged their outcomes together. How to split regions according to the shape of targets confuses the academic so that Region-CNN (RCNN) is put forward, which firstly combines ROI extraction and classification in one network as end-to-end learning by selective search. However, RCNN is made up of three models: selective search by multiple pixel-wise image scan, regional feature extraction and classification, each of which calls for demanding computing capability. For example, prediction for only one image generally requires 40-50 seconds, not to mention training process.

Thus, Fast RCNN integrates above three steps into one CNN model which is primarily responsible for selective search ROIs and the following fully connected layers are in charge of classification/regression within ROIs. Although prediction accelerates to around 2 seconds per image by Fast RCNN, selective search is still computationally expensive, which cannot handle with massive datasets. Consequently, Faster RCNN, the top of state-of-the-art object detection architectures, replaces selective search with RPN, for which this thesis exploits FRCNN as end-to-end framework.

Even if FRCNN is composed of two modules (demonstrated in fig. 3.24): RPN for ROI detection and Fast RCNN detector for classification, it is still a single and unified network [102]. After all, RPN serves as an attention mechanism to indicate which bounding boxes are salient enough in the feature maps extracted from the classifier backbone (convolutional layers in fig. 3.24 [102]). Taking the performance on benchmark datasets into consideration, our FRCNN adopts ResNet-50 as backbone initialized by the pre-trained ImageNet classification model [103].

With reference to RPN, fig. 3.25 precisely exposes how a series of convolutional operations denote the possibilities and locations of ROIs. The upper workflow contributes to target bounding box detection via classifying foreground and background, whereas the other calculates the offsets of bounding box regression for further location refinement. To be specific, after a 3×3 convolution for grouping local spatial information, classification branch creates anchor boxes for each pixel by a 1×1 convolutional operator at stride 1 with 18 filters. An anchor is centred at the sliding window in question with certain scale and aspect ratio (fig. 3.26) [102]. By default, our RPN follows the settings of [102] with 3 scales (16×8, 16×16, 16×32) and 3 aspect ratios (0.5, 1, 2) yielding 9 anchors at each sliding position, since ResNet-50 conducts 16 times downsampling as well.

In view of each anchor with two classes: foreground and background, the entire box-classification information can be stored in above 18 channels of 1×1 convolution. The criteria of binary label assignment for foreground and background also refer to [102]. An anchor whose IoU (explained in eq. (2.1)) is higher than 0.7 with any ground-truth box is designated as 1, while if an anchor's IoU ratio is lower than 0.3 for all ground-truth boxes, it hard codes as 0. Anchors without labels then do not engage in the training objective so that there won't be excessive anchors for training [102]. Similarly, bounding box regression branch is implemented by a 1×1 convolution at stride 1 with 36 filters, containing coordinates of top-left corner and bottom-right corner for each anchor. The transformations between target anchors and ground truths are learned as weights of 1×1 convolutional layer during training process, which addresses the extremely extra cost in ROI detection. As for proposal block, different with those benchmark datasets, such as PASCAL VOC, Microsoft COCO and so on, medical images for knees typically have merely one ROI, at most two. Accordingly, only top 2 anchors with the highest IoU are preserved but their IoU between each other should be less than an overlapping threshold, which is allocated as 0.7 in this thesis.

In order to unify the size of anchors for classifier, ROI pooling layer is built on the spatial pyramid concept. ROIs highlighted from RPN are equal to the size of original images, for which our ROI pooling layer foremost maps their size to feature maps with 1/16 spatial scale, followed by averagely division of width and height. The applied divisor is assigned as 7 in accordance with the initial research [104]. In each sub-block, max-pooling operation assists to fulfill the fixed-length output for fully connected layers, after which two branches establish again for both ROI refinement and knee OA classification. In spite of the complex excogitation, FRCNN expedites 10 times as Fast RCNN per image prediction, since the whole architecture can be trained synchronously without additional effort, likewise above traditional CNN-based approaches.

## 3.3 Decision visualization

The dramatic progress of ML in the form of deep neural networks has offered tremendous benefits with impressive results. However, different from those logical and symbolic reasoning approaches, they are perceived as "black box", a lack of internal functional understanding, which is the fatal weakness of automatic diagnosis. For the sake of stepping up knee OA quantification transparency, class-discriminating attention map visualization are appended to exhibit significant features for class assignment.

For a certain class, explaining where classification/regression takes into consideration for label determination amounts to collect weights of each feature map from the final convolutional layer. In this context, Zhao et al. presented CAM technique, which visualizes the weighted combination of the feature maps at the penultimate layer as heat-maps by a global average pooling [105]. Nevertheless, CAM has to retrain a linear classifier for each class, for which [106] subsequently came up with an efficient generalization of CAM, Grad-CAM. Instead of pooling, aiming at class $c$, Grad-CAM globally averages gradients of feature maps as weights depicted in eq. (3.10) [106].

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.10}$$

where $Z$ is the pixel number of each feature map. And the gradient of the score for class $c$ represents as $y^c$. Regarding $A_{ij}^k$, it denotes at (i,j) position the value of $k$th feature map. Having gathered relative weights, the coarse saliency map $L^c$ can be demonstrated as the weighted combination similar to CAM. Indicated in eq. (3.11), a ReLU activation function is employed to the linear combination of maps because only features that have a positive influence on the class are interested [106]. Those negative pixels are likely to belong to other categories in the image. As expected, without this ReLU, heat-maps, to a great extent, highlight more than the desired class [106].

$$L^c = ReLU(\sum_i \alpha_k^c A^k) \tag{3.11}$$

However, if an image contains multiple occurrences with slightly different orientations or views of the same class, several objects would fade away in the saliency map created by Grad-CAM. Moreover, merely parts of objects are spotlighted by Grad-CAM, due to its overlook of the significance disparity among pixels. Thus, Grad-CAM++ replaces global gradient average with a weighted average of the pixel-wise gradients [105]. With respect to how to explicitly code the structure of pixel weights, Grad-CAM++ reformulates eq. (3.10) to eq. (3.12) with the concept of Grad-CAM for reference [105]. The core of eq. (3.12) in question is how to express $\alpha_{ij}^{kc}$ with known symbols. Since the weights among pixels also contribute to the final classification score as eq. (3.13), rearranging the consolidation of eq. (3.12) and eq. (3.13), $\alpha_{ij}^{kc}$ can be declared as eq. (3.14) [105]. In case of confusion, here two iterators over the same activation map $A^k$, $(i, j)$ and $(a, b)$, are applied.

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU(\frac{\partial y^c}{\partial A_{ij}^k}) \tag{3.12}$$

$$y^c = \sum_k w_k^c \cdot \sum_i \sum_j A_{ij}^k \tag{3.13}$$

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \frac{\partial^3 y^c}{\{(\partial A_{ij}^k)^3\}}} \tag{3.14}$$

In reality, knee radiographs and MRIs rarely appear multi-targets. Furthermore, it is more reasonable to reveal the detail bits for determination rather than the entire joint. Considering the computing complexity, this research absorbs both Grad-CAM and Grad-CAM++ into evaluation candidates.

## 3.4 Model ensemble

Enlightened from the notion of ensemble learning, apart from aforementioned multimodality integration, the final module also takes on model fusion among preeminently trained models. Inspired by literature [8, 35, 75, 76], no matter whether the merged models are generated from the same training process or not, averaging their Softmax class posteriors is in common, which owns tremendous architecture compatibility. Taking the OA diagnosis context into account, it would be more reasonable to sort out the maximum score among predictions of top trained models. After all, the risk for one-more-grade diagnosis is much less than that of the inverse way. Spoken of in section 2.2.1, radiographic semi-quantitative criteria are surprisingly in vogue among automatic MRI prediction, which builds on a solid foundation for fusing multimodality predictions by above two operations. Since there is no previous reference to integrate multimodality diagnosis outcome as yet, both model fusion and multimodality integration assess the mean Softmax class posterior and the maximum among elite models regardless whether they are produced by the same architecture.

To sum up, each module of our multimodality based automatic knee OA quantification comprises a set of technique combinations (fig. 3.27), which would be hammered out based on profuse empirical evaluations revealed in chapter 4.

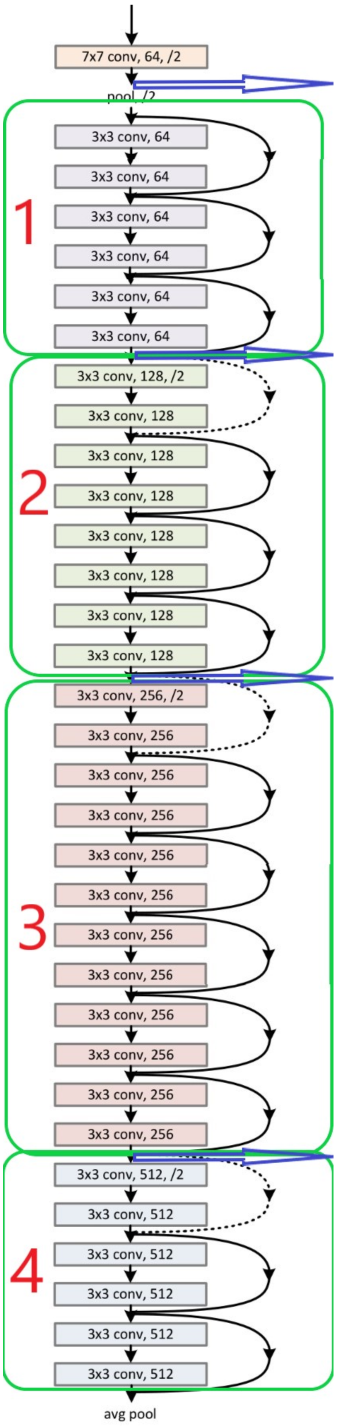
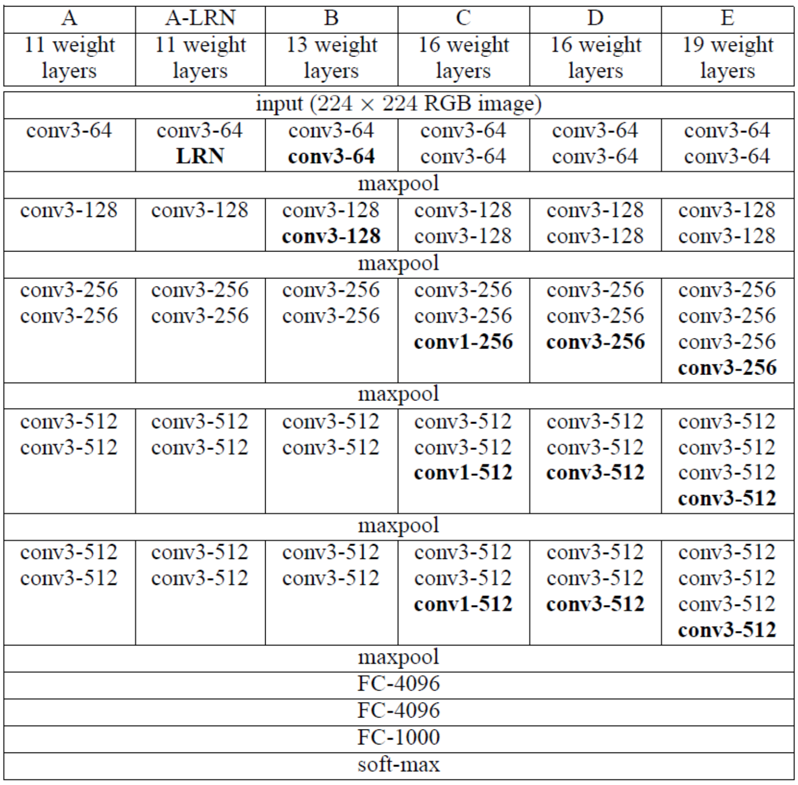


**Figure 3.18:** The contracting path of U-Net with the backbone ResNet-34



| A | A-LRN | B | C | D | E |
|---|---|---|---|---|---|
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 16 weight layers | 19 weight layers |
| input (224 × 224 RGB image) | | | | | |
| conv3-64 | conv3-64<br>**LRN** | conv3-64<br>**conv3-64** | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128<br>**conv3-128** | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 |
| maxpool | | | | | |
| conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256<br>**conv1-256** | conv3-256<br>conv3-256<br>**conv3-256** | conv3-256<br>conv3-256<br>conv3-256<br>**conv3-256** |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>**conv1-512** | conv3-512<br>conv3-512<br>**conv3-512** | conv3-512<br>conv3-512<br>conv3-512<br>**conv3-512** |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>**conv1-512** | conv3-512<br>conv3-512<br>**conv3-512** | conv3-512<br>conv3-512<br>conv3-512<br>**conv3-512** |
| maxpool | | | | | |
| FC-4096 | | | | | |
| FC-4096 | | | | | |
| FC-1000 | | | | | |
| soft-max | | | | | |

**Figure 3.19:** The original VGG architecture

The convolutional layer parameters are denoted as "receptive field size-number of channels", which are based on ILSVRC dataset as well as the input and output dimensions of VGG networks. LRN: Local Response Normalisation, a typical normalization approach in early stage CNNs. FC: Fully Connected Layer.

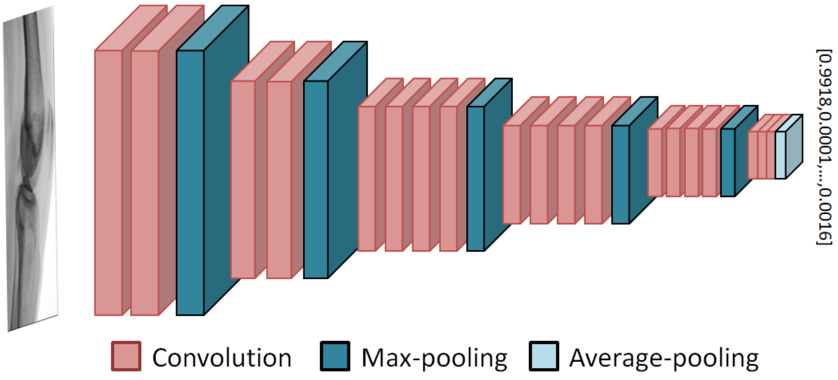


**Figure 3.20:** The revised VGG-19 architecture

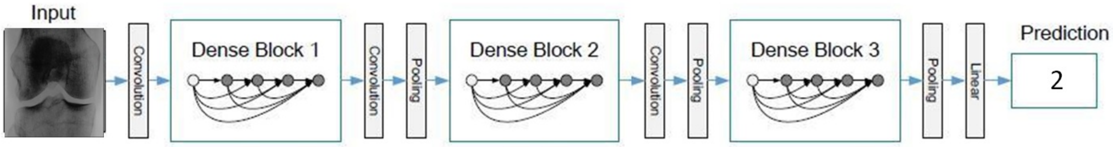
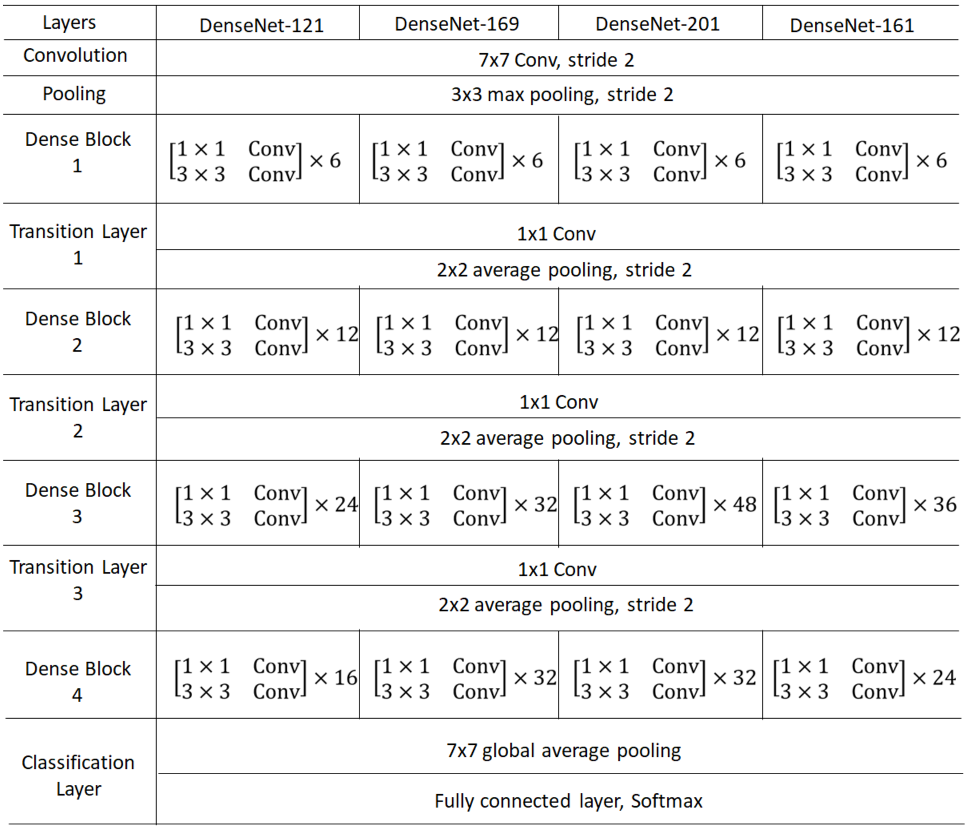


**Figure 3.21:** The workflow of DenseNet architecture



| Layers | DenseNet-121 | DenseNet-169 | DenseNet-201 | DenseNet-161 |
|---|---|---|---|---|
| Convolution | 7x7 Conv, stride 2 | | | |
| Pooling | 3x3 max pooling, stride 2 | | | |
| Dense Block 1 | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 6$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 6$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 6$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 6$ |
| Transition Layer 1 | 1x1 Conv | | | |
| | 2x2 average pooling, stride 2 | | | |
| Dense Block 2 | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 12$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 12$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 12$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 12$ |
| Transition Layer 2 | 1x1 Conv | | | |
| | 2x2 average pooling, stride 2 | | | |
| Dense Block 3 | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 24$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 32$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 48$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 36$ |
| Transition Layer 3 | 1x1 Conv | | | |
| | 2x2 average pooling, stride 2 | | | |
| Dense Block 4 | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 16$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 32$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 32$ | $\begin{bmatrix}1\times1 & \text{Conv}\\ 3\times3 & \text{Conv}\end{bmatrix}\times 24$ |
| Classification Layer | 7x7 global average pooling | | | |
| | Fully connected layer, Softmax | | | |

**Figure 3.22:** The mainstream DenseNet architectures

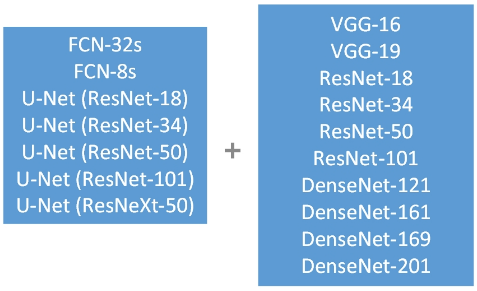


**Figure 3.23:** The design of CNN-based approaches

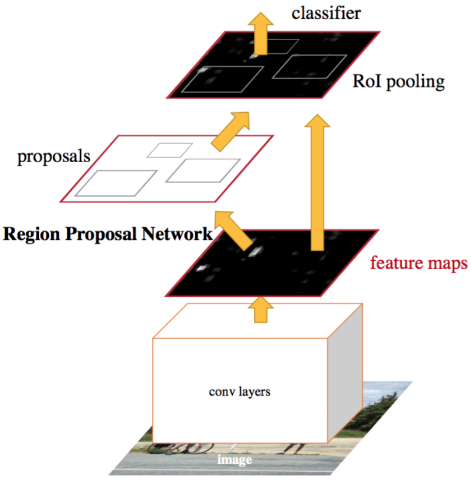


**Figure 3.24:** The FRCNN architecture

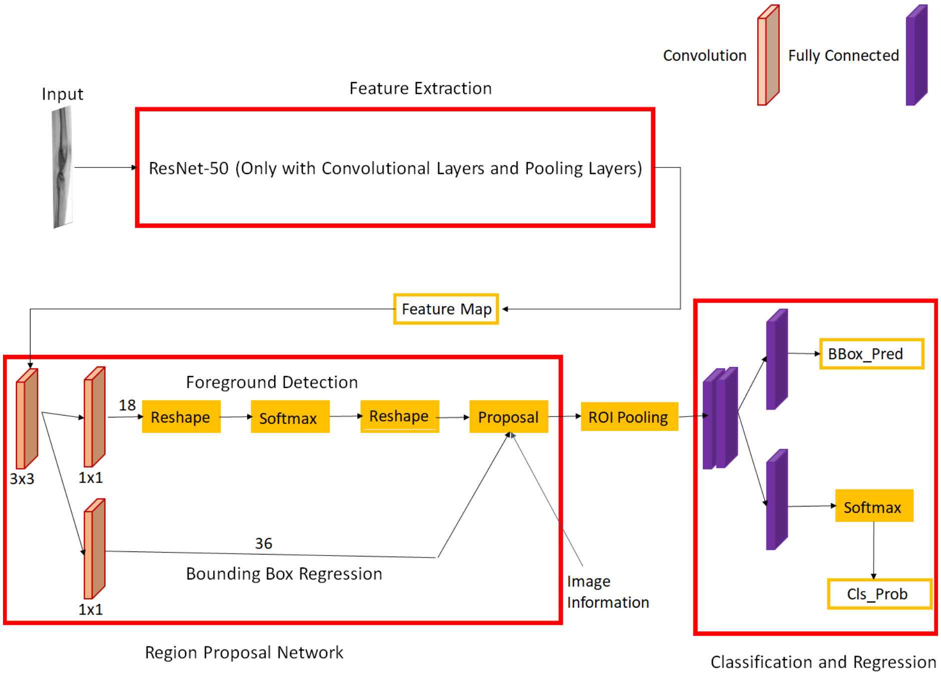


**Figure 3.25:** The workflow of FRCNN architecture

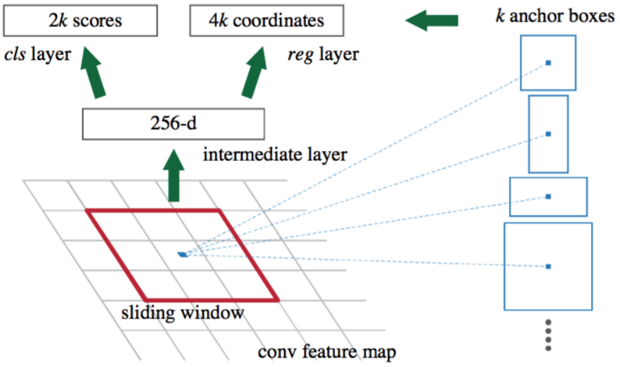


**Figure 3.26:** The anchor-based work principle

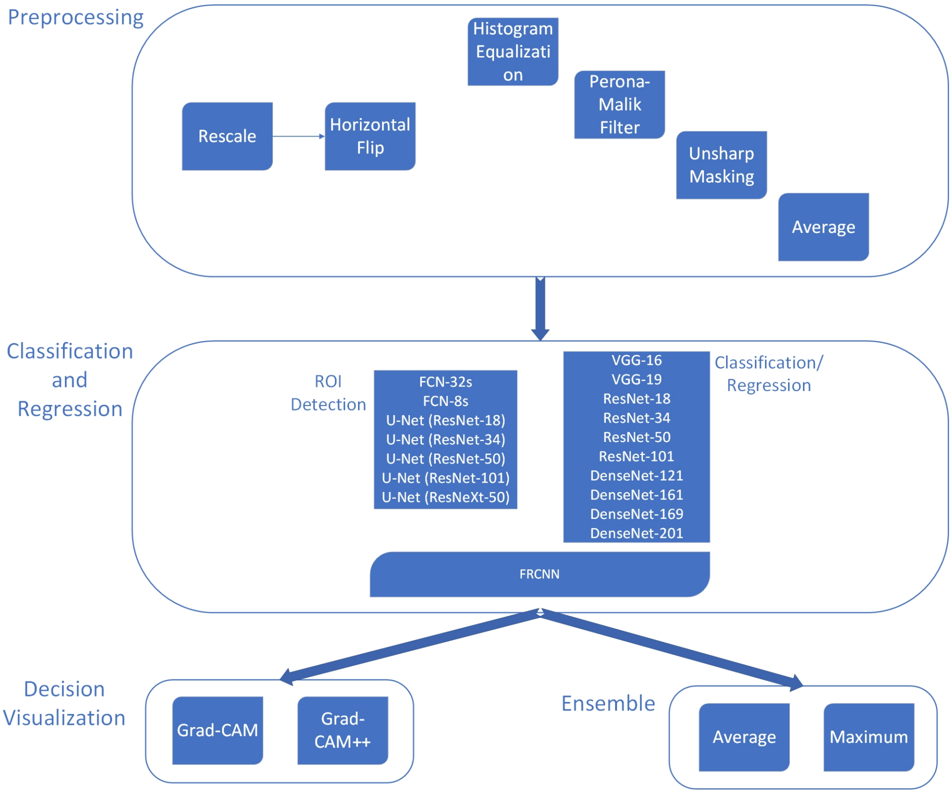


**Figure 3.27:** The technique to combine knee OA quantification system

# Chapter 4

# Evaluation

In order to break through the knee OA diagnosis bottleneck which stems from the defects of modality attributes, the core of this thesis establishes on the comparative quantification research among multi-perspectives of radiographs and MRIs in accordance to the scheme signified in fig. 3.27.

## 4.1 Experiment setup

Experiments were carried out on a machine having Intel(R) Xeon(R) CPU E5-2640, 256 of RAM, and Ubuntu 16.04 OS. The software stack consisting of Scikit-learn and Keras with the TensorFlow backend. The network training is carried out on an Nvidia GTX 1080i GPU with CUDA and cuDNN enabled to make the overall pipeline faster. For each hyperparameter group of the certain network structure, 5 repeated experiments are conducted, among which the best one is employed in the comparison.

### 4.1.1 Dataset

In consideration of the availability of public datasets, radiographic images and 2D MRI slices as well as their relative labels are rendered from MOST cohort, a standard public database for knee OA studies, which encompasses 3026 subjects and their six follow-up examinations in DICOM (Digital Imaging and Communications in Medicine) format [107]. In view of data integrity, our evaluation is based on their first visit data (V0). Although MOST possesses 3026 knee radiographic assessments, merely 2406 mutual patients are engaged in MRI collection. Due to the aim of multimodality integration, the following appraisal simply applies plain radiographs and MRI slices of above 2406 participants.

| Modality | Training/validation set | Test set |
|---|---|---|
| Radiograph (coronal plane) | 3409 | 1403 |
| Radiograph (sagittal plane) | 3345 | 1403 |
| MRI (axial plane) | 3273 | 1403 |
| MRI (sagittal localizer) | 3275 | 1403 |

**Table 4.1:** The distribution of the radiographs and MRIs in the training and test sets

| KL-0 | KL-1 | KL-2 | KL-3 | KL-4 |
|---|---|---|---|---|
| 2037 | 842 | 752 | 822 | 359 |

**Table 4.2:** KL scale distribution for grade 0-4 in MOST cohorts

Owing to information loss and view complementary, MRI slices from axial plane and localizer are opted from 4 perspectives (localizer, axial, sagittal and coronal). Hence, in total 4812 radiographs from coronal plane, 4748 radiographs from sagittal plane, 4676 MRI slices from axial plane and 4678 MRI slices from sagittal localizer contribute to the quantitative comparison among techniques picked in chapter 3. For the sake of deep learning training, each group of experiment images are randomly split into training/validation set (70% or so) and test set (30% or so), which maintains test sets with the same amount as Table 4.1 for fair comparisons.

As for labels, MOST proffers 3 types of OA semi-quantitative scores: KL scale, OARSI JSN progression gauged from medial tibiofemoral compartment and OARSI JSN progression gauged from lateral tibiofemoral compartment, whose detail distributions for each grade are respectively illustrated in Table 4.2 and Table 4.3. Considering the prevailing of KL in automatic MRI quantification and the demand of multimodality integration, it is feasible and indispensable to assign the same radiographic semi-quantitative labels for both radiographs and MRIs. Noticeably, JSN progressions of lateral tibiofemoral compartments are excessively imbalanced so that only the first two scoring metrics are adopted, not only for singling out the better evaluation criterion, but also for convincing the universality as well as the applicability of our proposed approach.

| Assessment position | JSN-0 | JSN-1 | JSN-2 | JSN-3 |
|---|---|---|---|---|
| Medial tibiofemoral | 2796 | 792 | 631 | 237 |
| Lateral tibiofemoral | 4093 | 165 | 149 | 68 |

**Table 4.3:** OARSI JSN progression distribution for grade 0-3 in MOST cohorts

### 4.1.2 Evaluation criteria

The entire automatic knee OA quantification design consists of both image segmentation and classification task, for which depending on different targets, metric selection is subdivided into two branches: ROI detection and classification/regression.

#### ROI detection

ROI detection can be regarded as 2-class segmentation with bounding boxes. Thus, IoU (mathematical explanation as eq. (2.1)) as the prime segmentation index is employed. After all, another leading metric, Dice coefficient (eq. (4.1)) would be always higher than IoU in the same situation theoretically.

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4.1}$$

In the above equation, $X$ is the ground truth and $Y$ is the prediction. When IoU is transferred into loss function of segmentation networks in eq. (4.1) can be transformed as shown in eq. (4.2).

$$IoU\,loss = 1 - \frac{X \cap Y}{X \cup Y} \tag{4.2}$$

IoU measures the detection precision from overlapping perspective by integrating the four bounds of a predicted box as a whole unit, while traditional loss functions in particular cross entropy series evaluate segmentation pixel-wisely. Based on information theory, entropy $\Delta I$ stands for the possible information quantity obtained from resources. If the possibility of certain event is $P_i$, $\Delta I = -ln(P_i)$. The BCE defined in eq. (4.3) is the optimal metric for distance between two sets in our ROI detection as 2-class bounding box prediction.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{4.3}$$

where $y_i$ is the ground truth for pixel $i$ and $p(y_i)$ represents the predicted probability that pixel $i$ is within ROIs. Certainly, $N$ is the number of pixels. However, the final averaging lead to a distinct shortcoming of BCE that every pixel owns the same standing for loss reduction. If the target barely takes up one-hundred of the whole picture, then the entire feature map is trend to be marked as background. Although it is common phenomenon for medical images, it is unlikely to annotate ROIs as such small area for MOST dataset. Hence, in order to take advantages of both methodologies, our loss function $L$ tots up BCE and IoU loss.

### Classification and regression analysis

In general, the acknowledged evaluation criterion for OA severity grading is accuracy, which is represented by the proportion of true positive and true negative in all evaluated cases [108], which can be stated mathematically as follows [108]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

where True Positive (TP) = the number of cases correctly identified as patient
False Positive (FP) = the number of cases incorrectly identified as patient
True Negative (TN) = the number of cases correctly identified as healthy
False Negative (FN) = the number of cases incorrectly identified as healthy

For the sake of imbalanced datasets, F-Measure and AUC are highly recommended as classification criteria rather than accuracy, which is extremely influenced by the dominated class [9]. Considering ROC is limited to the binary classification, F-Measure, together with precision and recall, is picked up in the thesis, shown as eq. (4.5), eq. (4.6) and eq. (4.7) correspondingly:

$$F - measure = \frac{(1 + \beta^2) Precision * Recall}{\beta^2 Precision + Recall} \tag{4.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

Aiming to OA diagnosis, precision and recall should be equally crucial so that as the symbol of the significance ratio between precision and recall, $\beta$ is fixed to 1. Then eq. (4.5) can be simplified as

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.8}$$

Taking the imbalanced class distribution into account, F1-score, precision and recall are calculated by macro-averaging, which treats the metric independently and equally for each class. Roughly, compared with micro-averages and weighted averages, it is the lowest but most impartial index. After all, class distributions between different scoring systems, perspectives and modalities vary in a broad range.

On the other hand, for an ordinal regression, a form of multiclass regression for which there is an inherent order between classes, the performance of knee OA prediction cannot be accounted for comprehensively by above metrics [109]. In order to take advantage of the precise error degree illustration by regression analysis, Gaudette and Japkowicz compare existing loss functions concluding that Mean Absolute Error (MAE) and MSE are the best so far [110].

$$MSE(Y) = E((Y - X)^2) \tag{4.9}$$

However, MAE lays its talent in the situation where the tolerance for small errors is lower, while MSE performs better when the severity of the errors is more vital [110]. Obviously, concerning OA diagnosis, one grade higher prediction is acceptable but large deviation takes too much risks. As a result, this thesis takes MSE as the loss function for training all the models and the main regression metric in eq. (4.9).

## 4.2 Preprocessing

As depicted in section 4.1.1, knee images of the whole MOST dataset are stored as DICOM files. Thereupon, during transforming DICOM files to PNGs (Portable Network Graphics), the black borders and the blank blocks are cut off, for which rescaling each group of medical images to their width/height minimums is requisite. MRI slices from axial plane are rescaled to 511 × 511 pixels, while those from localizer are in the shape of 255 × 255 pixels. And radiographs from sagittal plane are set to 1520 × 2047 pixels (Width × Height). As for radiographs from coronal plane, due to the setting as bilateral PA fixed flexion knee X-Ray images, in order to split both knees, they are cut apart from the middle of width and adjusted to 1023× 2047 pixels.

After horizontal flipping all the right knees to left, radiographs conduct histogram equalization as analysed in section 3.1, followed by diversified combinations among Perona-Malik filtering and unsharp masking. On the other hand, before multi-slice averaging, MRIs execute varied groups of Perona-Malik filtering and unsharp masking. In terms of the instruction of the picked Python library, MedPy, the possible $K$ of Perona-Malik filtering are selected as 20 (recommended minimum setting) and 50 (default setting), whereas iterations are assigned to 1 and 5. With regards to MRI slice integration, having discarded the first and the last one-third slices of each series, the middle part is averaged per pixel as control trail with three specified MRI slices: the median and its two adjacent slices.

As preliminary experiment, the lightweight VGG-19 is trained from sketch with 32 samples per batch to minimize MSE by the default Adaptive moment estimation (Adam) optimizer for a fast but accurate comparison among the following radiographic preprocessing approaches expressed in Table 4.4. The accuracy of test set is monitored by EarlyStopping callback function during training process. If validation accuracy does not rise after 250 epochs, the training process would be halted.

Before we start training the neural networks, image standardization is done in which

| Approach | Filtering iterations | $K$ for filter | Unsharp masking kernel |
|---|---|---|---|
| 1 | 1 | 50 | —— |
| 2 | 1 | 50 | SHARPEN |
| 3 | 1 | 50 | EDGE_ENHANCE |
| 4 | 5 | 20 | —— |
| 5 | 5 | 20 | SHARPEN |
| 6 | 5 | 20 | EDGE_ENHANCE |
| 7 | 5 | 50 | —— |

**Table 4.4:** Preprocessing approaches for radiographs

the mean pixel value is substracted from each pixel and dividing the difference by the standard deviation of pixel values. Compared to standardizing each image separately as sample-wise standardization, feature-wise standardization calculated on the whole datasets is adopted for both training/validation sets and test sets, followed by image normalization which traditionally rescales pixel values into [0,1] by pixel-wisely multiplying with 1/255 ratio. After all, the full image sets merely consist of grey-scale images, whose pixel value range is between 0 and 255.

In case ROI detection step has effect on the comparison, pre-experiments work with the manually annotated ROIs of radiographs. Due to the flexible sizes of labelled knee joints, radiographic ROIs from sagittal plane are resized to 352 $\times$ 544 pixels, while those captured from coronal plane contain 320 $\times$ 352 pixels.

Pursuant to the original VGG-19 architecture illustrated in fig. 3.19, our training is always stuck in the local optimal solution. Having observed plenty of trails on receptive field sizes, filter numbers of each convolutional layer and hidden nodes numbers of each fully connected layer, [3×3] kernel size is preserved. Since our image components are not complicated, redundant feature maps are supposed to results in overfitting and trapping validation performance. Thus, filter numbers of our VGG-19 shrink 32 times per layer (Stack 1: 2, Stack 2: 4, Stack 3: 8, Stack 4: 16, Stack 5: 16). With respect to hidden nodes numbers of the first two fully connected layers, in consideration of the difference between our dataset capability and the image quantity of ILSVRC, they are diminished to 128 and 32. As for regression analysis, it is pretrained by the weights of corresponding trained classifier.

On the basis of our more lightweight VGG-19, Table 4.5 and Table 4.6 give an all-round illustration that Approach 1 (Histogram Equalization + Slight Perona-Malik Filter) is ideal for X-Ray images. After one iteration Perona-Malik filtering with $K$ as high as 50, the processed images are nearly the same with the original one, which extends the conclusion in [14] to radiographs that Perona-Malik filtering does not contribute more to MRI enhancement on the basis of histogram equalization. The more iterations indeed smooths the bone shapes, which results in the worse cases, likewise blurred images generated by

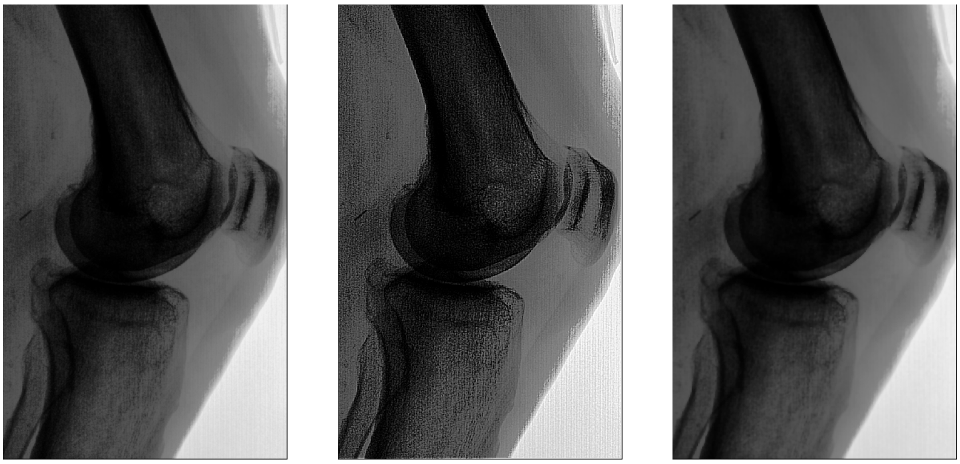
| Metric | Approach | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| JSN | 1 | 0.8176 | 0.76 | 0.70 | 0.73 | 0.0779 |
| JSN | 2 | 0.6805 | 0.68 | 0.67 | 0.67 | 0.1346 |
| JSN | 3 | 0.6980 | 0.7 | 0.67 | 0.68 | 0.1273 |
| JSN | 4 | 0.7847 | 0.68 | 0.56 | 0.59 | 0.1019 |
| JSN | 5 | 0.7917 | 0.71 | 0.63 | 0.65 | 0.0959 |
| JSN | 6 | 0.7500 | 0.66 | 0.58 | 0.60 | 0.1106 |
| JSN | 7 | 0.8125 | 0.71 | 0.62 | 0.65 | 0.0881 |
| KL | 1 | 0.6712 | 0.64 | 0.59 | 0.60 | 0.0924 |
| KL | 2 | 0.5773 | 0.59 | 0.57 | 0.57 | 0.1464 |
| KL | 3 | 0.5311 | 0.55 | 0.51 | 0.52 | 0.1414 |
| KL | 4 | 0.6320 | 0.53 | 0.53 | 0.50 | 0.1149 |
| KL | 5 | 0.6280 | 0.55 | 0.53 | 0.52 | 0.1298 |
| KL | 6 | 0.6000 | 0.56 | 0.50 | 0.52 | 0.1161 |
| KL | 7 | 0.6520 | 0.62 | 0.58 | 0.59 | 0.1222 |

**Table 4.5:** Classification and regression results for radiographic approaches (coronal plane)

| Metric | Approach | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| JSN | 1 | 0.6908 | 0.47 | 0.43 | 0.42 | 0.1117 |
| JSN | 2 | 0.4289 | 0.45 | 0.38 | 0.38 | 0.2446 |
| JSN | 3 | 0.4201 | 0.42 | 0.40 | 0.40 | 0.2467 |
| JSN | 4 | 0.4792 | 0.37 | 0.32 | 0.32 | 0.2354 |
| JSN | 5 | 0.4583 | 0.38 | 0.36 | 0.34 | 0.2524 |
| JSN | 6 | 0.4375 | 0.52 | 0.32 | 0.32 | 0.2403 |
| JSN | 7 | 0.4653 | 0.41 | 0.31 | 0.31 | 0.2322 |
| KL | 1 | 0.5544 | 0.46 | 0.42 | 0.40 | 0.1192 |
| KL | 2 | 0.4156 | 0.44 | 0.39 | 0.39 | 0.1557 |
| KL | 3 | 0.3517 | 0.36 | 0.33 | 0.32 | 0.2075 |
| KL | 4 | 0.4480 | 0.36 | 0.34 | 0.34 | 0.1797 |
| KL | 5 | 0.4400 | 0.39 | 0.37 | 0.37 | 0.1872 |
| KL | 6 | 0.4080 | 0.32 | 0.31 | 0.30 | 0.2220 |
| KL | 7 | 0.4520 | 0.40 | 0.38 | 0.39 | 0.1909 |

**Table 4.6:** Classification and regression results for radiographic approaches (sagittal plane)

Approach 7 (as displayed in fig. 4.1(c)). As for the lower $K$, Approach 4-6 keep more noises, for which not only they have higher misclassification rates, but also their image qualities are not enriched by edge sharpening. After all, those noises are also enhanced at the same time. Similarly, having transformed by global contrast enhancement (histogram equalization), the high-resolution radiographs in Approach 2 and 3 have already hold strong contrast so that unsharp masking even brings more noises into, as indicated in fig. 4.1(a) and fig. 4.1(b).



(a) Approach 1 (b) Approach 3 (c) Approach 7

**Figure 4.1:** Radiographs after different preprocessing

Drawing a salutary lesson from radiograph preprocessing, $K$ in Perona-Malik filters is assigned as 50 for MRI preprocessing. Since the edges of MRIs are much blurrier than those of X-Ray images, there is no need to lift up the default $K$. With respect to iterations, it retains as 1, given the fact that multi-slice averaging has smoothed the acquired images. Owing to the lack of global contrast enhancement, edge sharpening comparison is remained as Table 4.7. In order to corroborate the excellence of slice averaging, the unsharp masking kernel is specially picked as EDGE_ENHANCE, the kernel with marginally poorer performance in radiographs.

Distinct with X-Ray images, now that the whole MRIs are small enough for neural networks and artefacts inside are within acceptable interval, they are directly trained pursuing the architecture, lightweight VGG-19, and the training settings of the radiograph preprocessing evaluation. Within expectation, averaging is the optimum approach for multi-slice integration (seen in Table 4.8 and Table 4.9). As a result of short MRI sequences, in particular those from localizers, the median slice also achieves approximately similar F1-score, followed by the slice after median. This depends on the median slice determination method that if there are even slices in total, the previous one is handled. Under this condition, the slice prior to median remains at the last places. Synthesizing the corresponding entries of Table 4.8 and Table 4.9 (discarding Approach 1 with slice aver-

| Approach | Unsharp masking kernel | Slice selection |
|---|---|---|
| 1 | EDGE_ENHANCE | Average |
| 2 | EDGE_ENHANCE | Median |
| 3 | EDGE_ENHANCE | One slice prior to median |
| 4 | EDGE_ENHANCE | One slice after median |
| 5 | SHARPEN | Median |
| 6 | SHARPEN | One slice prior to median |
| 7 | SHARPEN | One slice after median |

**Table 4.7:** Preprocessing approaches for MRIs

age), it turns out a flagrant contrast between two edge sharpening kernels in Table 4.10. Opposed to radiographs, EDGE_ENHANCE accomplishes lightly higher F1-score and lower MSE than SHARPEN, for which Approach 1 maintains its unsharp mask kernel in further assessments.

To sum up, both X-Ray images and MRIs would be pre-processed as their Approach 1 (Radiograph: Histogram Equalization + 1 iteration Perona-Malik Filtering with $K = 50$; MRI: 1 iteration Perona-Malik Filtering with $K = 50$ + Unsharp Masking with EDGE_ENHANCE Kernel + Multi-slice Averaging). Their accuracies, even the majority of precisions, are notably higher than others. However, restricted to the imbalanced dataset, their recalls cannot remain at the same level, which exhibits the considerable promotion potential, in particular with JSN grading system. On the whole, by reason of the irregular class distribution, the trained models always stagnate at the local optimal solution, only concentrating on classes rich in data, which leads to the unsatisfactory results and the sharp decline from accuracy to precision, especially for sagittal radiographs and MRIs.

However, if weighted average evaluation metrics are taken into account, their values would soar by 10% or so, even over 15% for JSN grading system. For example, the weighted average precision, recall and F1-score of Approach 1 for axial MRIs under JSN scores are respectively 0.68, 0.66 and 0.57, while those under KL grades are 0.44, 0.49 and 0.40. Although X-Ray images from coronal plane can handle the imbalanced data distribution superbly with an impressive performance, which is similar to state-of-art top researches ([8][28][36]), it still possesses promising prospect for progress. Furthermore, for the sake of multimodality integration, balancing dataset is undoubtedly imperative.

On this occasion, the classification and regression performance based on JSN scores is distinctly superior rather than that of KL, likewise Table 4.7, whereupon further experiments lay more emphasis on classification/regression under OARSI JSN semi-quantitative system. Having nominating final architecture candidates on the basis of JSN, our approach would be ascertained by KL grades moreover.

| Metric | Approach | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| JSN | 1 | 0.6662 | 0.40 | 0.28 | 0.34 | 0.1244 |
| JSN | 2 | 0.3964 | 0.35 | 0.35 | 0.34 | 0.1438 |
| JSN | 3 | 0.3417 | 0.25 | 0.29 | 0.25 | 0.1831 |
| JSN | 4 | 0.5224 | 0.27 | 0.31 | 0.28 | 0.1667 |
| JSN | 5 | 0.5224 | 0.27 | 0.31 | 0.27 | 0.1574 |
| JSN | 6 | 0.4403 | 0.21 | 0.26 | 0.22 | 0.1601 |
| JSN | 7 | 0.5299 | 0.27 | 0.32 | 0.29 | 0.1853 |
| KL | 1 | 0.4774 | 0.38 | 0.37 | 0.36 | 0.1382 |
| KL | 2 | 0.3179 | 0.36 | 0.29 | 0.27 | 0.1615 |
| KL | 3 | 0.2976 | 0.28 | 0.28 | 0.27 | 0.2307 |
| KL | 4 | 0.4195 | 0.32 | 0.33 | 0.32 | 0.1953 |
| KL | 5 | 0.4322 | 0.33 | 0.26 | 0.29 | 0.2130 |
| KL | 6 | 0.3602 | 0.31 | 0.31 | 0.30 | 0.2270 |
| KL | 7 | 0.4195 | 0.37 | 0.31 | 0.29 | 0.1624 |

**Table 4.8:** Classification and regression results for MRI approaches (sagittal localizer)

| Metric | Approach | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| JSN | 1 | 0.6603 | 0.58 | 0.31 | 0.40 | 0.1232 |
| JSN | 2 | 0.3690 | 0.37 | 0.38 | 0.36 | 0.1544 |
| JSN | 3 | 0.3303 | 0.32 | 0.30 | 0.30 | 0.1844 |
| JSN | 4 | 0.5149 | 0.31 | 0.30 | 0.26 | 0.1568 |
| JSN | 5 | 0.5000 | 0.25 | 0.30 | 0.27 | 0.1621 |
| JSN | 6 | 0.4962 | 0.25 | 0.29 | 0.27 | 0.1638 |
| JSN | 7 | 0.4851 | 0.25 | 0.29 | 0.26 | 0.1616 |
| KL | 1 | 0.4943 | 0.38 | 0.29 | 0.33 | 0.1321 |
| KL | 2 | 0.4156 | 0.31 | 0.31 | 0.31 | 0.1400 |
| KL | 3 | 0.2791 | 0.23 | 0.22 | 0.13 | 0.1596 |
| KL | 4 | 0.3856 | 0.36 | 0.33 | 0.33 | 0.2003 |
| KL | 5 | 0.3517 | 0.26 | 0.28 | 0.25 | 0.1542 |
| KL | 6 | 0.3447 | 0.27 | 0.26 | 0.25 | 0.1862 |
| KL | 7 | 0.3856 | 0.24 | 0.29 | 0.25 | 0.1552 |

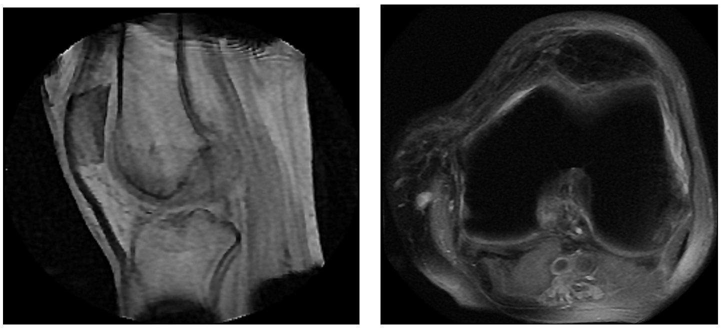
**Table 4.9:** Classification and regression results for MRI approaches (axial plane)

| Kernel | View | Label | F1-score | MSE |
|---|---|---|---|---|
| EDGE_ENHANCE | Localizer | JSN | 0.29 | 0.1645 |
| EDGE_ENHANCE | Localizer | KL | 0.29 | 0.1958 |
| EDGE_ENHANCE | Axial | JSN | 0.31 | 0.1651 |
| EDGE_ENHANCE | Axial | KL | 0.26 | 0.1666 |
| SHARPEN | Localizer | JSN | 0.25 | 0.1676 |
| SHARPEN | Localizer | KL | 0.29 | 0.2008 |
| SHARPEN | Axial | JSN | 0.27 | 0.1625 |
| SHARPEN | Axial | KL | 0.25 | 0.1652 |

**Table 4.10:** Unsharp masking kernel comparison among MRIs

## 4.3 ROI detection

As stated in section 4.2, without discernible artifacts, knee joints clearly stand out against the black background in the preprocessed MRIs as shown in fig. 4.2 in which there is an an appropriate scale for direct neural network training. Therefore, MRIs from both perspectives in the MOST cohorts are not required to extract ROIs prior to classifiers and consequently we focus only on radiographs.



(a) From sagittal localizer (b) From axial plane

**Figure 4.2:** Preprocessed MRIs

### 4.3.1 CNN-based approach

Served as a vital stage of traditional CNN-based approach described in section 3.2.1, pure ROI detection highly depends on the labelled ground truths for the segmentation training. Thereupon, the knee joints are manually annotated by a Python-based image annotation tool, LabelImg, without size limitation, since the shapes of knee joints from

| Architecture | Batch size | IoU score | Loss |
|---|---|---|---|
| FCN-32s | 16 | 0.8262 | 0.2374 |
| FCN-8s | 64 | 0.8444 | 0.2237 |
| U-Net (ResNet-18) | 80 | 0.9807 | 0.0257 |
| U-Net (ResNet-34) | 64 | 0.9430 | 0.1075 |
| U-Net (ResNet-50) | 32 | 0.9429 | 0.0774 |
| U-Net (ResNet-101) | 24 | 0.9370 | 0.0737 |
| U-Net (ResNeXt-50) | 24 | 0.9361 | 0.0788 |

**Table 4.11:** ROI detection results of radiographs from coronal plane

| Architecture | Batch size | IoU score | Loss |
|---|---|---|---|
| FCN-32s | 32 | 0.7657 | 0.3312 |
| FCN-8s | 64 | 0.7725 | 0.3320 |
| U-Net (ResNet-18) | 96 | 0.9637 | 0.0729 |
| U-Net (ResNet-34) | 64 | 0.8466 | 0.2732 |
| U-Net (ResNet-50) | 64 | 0.9184 | 0.1506 |
| U-Net (ResNet-101) | 24 | 0.8940 | 0.1519 |
| U-Net (ResNeXt-50) | 32 | 0.8964 | 0.1497 |

**Table 4.12:** ROI detection results of radiographs from sagittal plane

sagittal plane are altered flexibly. According to the coordinates of bounding boxes, the ground truths of X-Ray images are generated to binary images with the same sizes of the input images (Sagittal: 1520 × 2047 pixels; Coronal: 1023× 2047 pixels).

Depict in section 4.1.2, this thesis upgrades previous segmentation loss function from single BCE or IoU loss to their sum so that after input normalization, FCN and U-Net can be thoroughly trained by the Adam optimizer with default parameters and straightforwardly monitored by EarlyStopping callback function via IoU score of the test dataset. After all, the combination of BCE loss and accuracy only measures how many percent of input images whose IoU score are higher than 50%, for which the highest accuracy does not stand for the most outstanding performance. Regarding the group of IoU loss and IoU score, it is void of pixel-wise perspective.

Having observed the growth tendency in several trails, the EarlyStopping patience of FCN is specified as 200, whereas that of U-Net is scheduled to 80. As for another hyperparameter, batch size, it is tuned in the learning procedure to explore the prefect settings, whose outcomes are presented in Table 4.11 and Table 4.12.

As the template of our baseline FCN-32s, the mean IoU for MOST cohort in [28] reached 0.81, which is quite approximate our baseline IoU average (0.8262). Evidently, this confirms the rationality and comparability for our evaluation. In spite of the fact that the architecture upgrade from FCN-32s to FCN-8s truly promotes the IoU score, the increment around 0.01 for X-Ray images acquired from both planes is far away from expectation. Since before settling on the final network, [28] had already verified plentiful convolution stages, the number of filters and kernel sizes in each convolution layer, our experiments does not try those parameters out. Compared with IoUs calculated from U-Net segmentation, the least 0.1 addition exposes the fateful factor behind the insufficient FCN performance, the defects of architecture. After all, concatenation for each pooling layer impeccably preserves all the local features, while vast spatial information is still lost during downsampling and adding in FCN.

Concerning U-Net backbones, ResNet-18 visibly has a brilliant achievement no matter which perspective of X-Ray images and which evaluation criterion is applied. Rather than 98.3% accuracy in [35], which represents ROIs in 98.3% of validated coronal radiographs are detected with IoU>0.5, the mean IoU of our U-Net with ResNet-18 architecture is 0.9807, whose percentages of detected ROIs with IoU>0.5 and IoU>0.75 are 100%. Considering knee joints from coronal plane are more monotonous than those from sagittal plane, average IoUs of sagittal X-Ray images are 5% or so lower. In the meanwhile, U-Net with other backbones also behave superb enough. However, aiming at 2-class bounding box detection, a relatively simple segmentation task, U-Net with more complicated structures are more likely overfitted, which is substantiated by a mild decrease trend on IoU scores in Table 4.11. This also illustrates why GANs are not applied in our ROI detection. After all, the powerful GAN is overqualified for binary bounding box detection. At the same time, this overfitting issue also explains why backbones with more layers are more suitable for smaller batch sizes and aiming at ROI detection for sagittal radiographs, expanded batch sizes is called for.

In conclusion, U-Net with complex backbones are more suitable for sophisticated segmentation tasks, whereas U-Net based on ResNet-18 is picked for our ROI detection of traditional CNN.

### 4.3.2 RPN

Distinguished from pure ROI detection, RPN of FRCNN is trained based on feature maps and bounding box coordinates so that the ROI coordinates produced by LabelImg remain to be employed but cannot be the only decisive factor any more. Apart from vertex coordinate distances, the Adam optimizer has to optimize classification BCE on the proposed ROIs as well. In order to shrink convergence duration, pre-trained ResNet-50 weights from Keras are loaded as initial parameters. Since whether the entire training stops or not is chiefly determined by classification accuracy at the last step, this subsection takes models trained after 30 epochs to calculate IoUs for comparison (seen in Table 4.13). Thanks to 1000 steps per epoch, which is mainly designed for OA grading, in our case, generally after 30 epochs, the regression loss computed by coordinate distances between proposed

| Approach | Perspective | IoU≥0.25 | IoU≥0.5 | IoU≥0.75 | IoU score |
|---|---|---|---|---|---|
| FCN in [28] | Coronal | 99.5% | 98.4% | 85.0% | 0.81 |
| U-Net (ResNet-18) | Coronal | 100.0% | 100.0% | 100.0% | 0.9807 |
| RPN | Coronal | 100.0% | 100.0% | 100.0% | 0.9899 |
| U-Net (ResNet-18) | Sagittal | 100.0% | 100.0% | 100.0% | 0.9637 |
| RPN | Sagittal | 100.0% | 100.0% | 100.0% | 0.9297 |

**Table 4.13:** Comparison between IoUs of RPNs and pure ROI detection approaches

and target bounding boxes would descends lower than 0.01 and nearly stop to revise.

Obviously, ROIs proposed by RPNs are more accurate than previous papers. Nevertheless, its performance is regulated by the final classification outcomes. The more surpassing radiographs are classified, the more likely RPNs are fostered so that it can attain similar even better IoU score than U-Net for coronal radiographs, while respecting sagittal radiographs, RPN is limited a bit by the grading difficulty. In practice, RPN owns more apparent operation simplicity because it not only merges two model training to skip the ROI extraction step in macroscopic view, but also ignores ground truths from detail implementation directly replaced by bounding box coordinates, which leaves prediction contour detection out. Hence, under synthetic consideration, the predicted bounding boxes from RPNs are straightly extracted for all the following experiments, although RPN's IoU average for sagittal X-Ray images is slightly lower than that of U-Net equipped with ResNet-18, which still can take up the second ranking with superior performance.

## 4.4 Classification and regression analysis

Having gained experiences from pre-experiments, previous to classification and regression analysis, in terms of JSN scores, dataset balancing is conducted via image augmentation for each perspective of plain radiographs and MRIs to guarantee the sample amount balance among classes. According to Table 4.3, triple extra images of grade 1 and grade 2 are demanded, while elevenfold additional images of grade 3 are required. Inspired by [8, 35], every image of grade 1 and grade 2 is rotated 5°, 355° and 10° clockwise, since diverse angles creates higher variability within datasets. As for grade 3, 230 images per perspective of each modality from other visits (V2, V3 and V5) are randomly selected for following preprocessing steps as Approach 1.

Having automatically extracted radiographic ROIs from RPN prediction, all the grade 3 knee joint images are together rotated 5°, 10° and 15° clockwise as well as counter-clockwise. In case one subject rotated in different angles appears in the training/validation dataset and the test dataset at the same time, the rotation procedure and

| Modality | Perspective | JSN=0 | JSN=1 | JSN=2 | JSN=3 |
|---|---|---|---|---|---|
| Radiograph | Coronal | 862 | 868 | 660 | 924 |
| MRI | Axial | 862 | 868 | 660 | 924 |
| Radiograph | Sagittal | 862 | 868 | 660 | 816 |
| MRI | Localizer | 862 | 868 | 660 | 818 |

**Table 4.14:** Class distributions under JSN scores in test sets

| Optimizer | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| Adam | 0.8633 | 0.86 | 0.86 | 0.85 | 0.0535 |
| SGD | 0.8896 | 0.88 | 0.88 | 0.88 | 0.0451 |

**Table 4.15:** Comparison between Adam and SGD optimizer based on VGG-19 with the balanced coronal radiographs

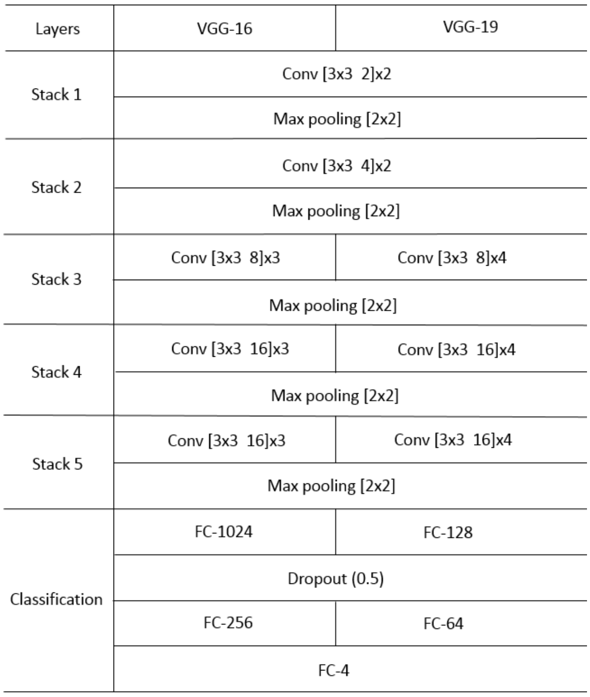
the extra image import are carried out in the training/validation sets and test sets separately. Thereupon, with the growing need of model adaptability, their class distributions achieve almost balanced as Table 4.14. On the balanced datasets, radiographic ROIs extracted from RPNs are resized to their width/height minimums: 336 × 359 pixels for coronal X-Ray images and 355 × 568 pixels for sagittal radiographs. Besides, on account of possible GPU memory limitation, especially for networks over 150 layers, MRIs from axial plane are rescaled to 360 × 360 pixels on the whole.

As what pre-experiments have done, images on the test datasets are feature-wisely standardized and centred with the same distribution of the training/validation images. Prior to model training, image normalization is fulfilled based on dividing pixel values by 255. Differing from preliminary experiments, this section cares more about functional performance, such as accuracy, precision, recall and F1-score, rather than convergence speed. Accordingly, the Adam optimizer is replaced by a more reliable optimizer, SGD, to lessen MSE. SGD is shorten from Stochastic Gradient Descent, however, the mini-batch import modifies it to mini-batch gradient descent in this thesis. Even though SGD is more strict with initialization and parameter settings, having attempted innumerable combinations of learning rates, initial weights, weight decays, momentums and Nesterov based on hyperparameter optimization with Bayesian algorithm implemented by Sherpa, our final SGD settings (Learning rate: 0.01, Weight decays: 1e-6, Momentum: 0.95, Nesterov: True, Initializer: He_normal, Regularizer: L2) can realize more convincing performance than that of Adam, which can be demonstrated by Table 4.15.

### 4.4.1 VGG architectures

Thanks to the expertises acquired from abundant pre-experiments on batch sizes, VGG networks are trained from sketch still with 32 samples per batch. Regarding the number of epochs, it depends on the validation accuracy monitored by EarlyStopping callback function in the training process. In order to make sure the highest accuracy can be captured and avoid local optimal solutions, the patience of EarlyStopping function is set to 300 and the minimum increment remains at 0.

In view of the image monotony, VGG-16 and VGG-19 persist our modification for convolutional filter numbers. Due to the dataset expansion, Dropout is imported for overfitting prevention. Concerning Dropout rates, along with numbers of fully connected layers and hidden node numbers per fully connected layer, ahead of eventual architecture determination, ample attempts are succeeded in. The detailed final network settings are noted in fig. 4.3. Indeed, instead of excessively extra hidden nodes, the image set augmentation demands more on moderate Dropout rate, while the quality differences among extracted feature maps caused by lacking convolutional layers have to be remedied by more hidden nodes in fully connected layers. Even if VGG-16 enlarges



<table>
<tr><th>Layers</th><th>VGG-16</th><th>VGG-19</th></tr>
<tr><td rowspan="2">Stack 1</td><td colspan="2">Conv [3x3 2]x2</td></tr>
<tr><td colspan="2">Max pooling [2x2]</td></tr>
<tr><td rowspan="2">Stack 2</td><td colspan="2">Conv [3x3 4]x2</td></tr>
<tr><td colspan="2">Max pooling [2x2]</td></tr>
<tr><td rowspan="2">Stack 3</td><td>Conv [3x3 8]x3</td><td>Conv [3x3 8]x4</td></tr>
<tr><td colspan="2">Max pooling [2x2]</td></tr>
<tr><td rowspan="2">Stack 4</td><td>Conv [3x3 16]x3</td><td>Conv [3x3 16]x4</td></tr>
<tr><td colspan="2">Max pooling [2x2]</td></tr>
<tr><td rowspan="2">Stack 5</td><td>Conv [3x3 16]x3</td><td>Conv [3x3 16]x4</td></tr>
<tr><td colspan="2">Max pooling [2x2]</td></tr>
<tr><td rowspan="4">Classification</td><td>FC-1024</td><td>FC-128</td></tr>
<tr><td colspan="2">Dropout (0.5)</td></tr>
<tr><td>FC-256</td><td>FC-64</td></tr>
<tr><td colspan="2">FC-4</td></tr>
</table>

**Figure 4.3:** Final architectures of VGG-16 and VGG-19

The convolutional layer parameters are denoted as receptive field size and number of channels. FC: Fully Connected Layer.

| Network | Perspective | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| VGG-16 | Coronal | 0.8739 | 0.87 | 0.87 | 0.87 | 0.0488 |
| VGG-19 | Coronal | 0.8896 | 0.88 | 0.88 | 0.88 | 0.0451 |
| VGG-16 | Sagittal | 0.7130 | 0.70 | 0.70 | 0.70 | 0.1048 |
| VGG-19 | Sagittal | 0.7489 | 0.75 | 0.74 | 0.74 | 0.1019 |

**Table 4.16:** VGG classification and regression performance on radiographs

| Network | Perspective | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| VGG-16 | Sagittal | 0.6742 | 0.67 | 0.66 | 0.65 | 0.1155 |
| VGG-19 | Sagittal | 0.6835 | 0.69 | 0.67 | 0.66 | 0.1096 |
| VGG-16 | Axial | 0.6307 | 0.58 | 0.60 | 0.58 | 0.1246 |
| VGG-19 | Axial | 0.6261 | 0.59 | 0.60 | 0.59 | 0.1316 |

**Table 4.17:** VGG classification and regression performance on MRIs

octuple and quadruple numbers of hidden nodes for each layer, the diagnosis from VGG-19 is more reliable overall (Table 4.16 and Table 4.17). This excellence is more significant for sagittal knee joints, since with more formations, classifiers need more fitting feature maps, which relies on convolutional layers. As for MRIs from axial plane, appropriate JSN-related features are extracted more challengingly, which gives rise to this similar performance. Therefore, the dense evaluation concept mentioned in section 3.2.1 is developed on the basis of VGG-19 to import more convolutional layers for feature extraction. On this occasion, hyperparameter optimization is conducted for culling the most suitable filter number of additional convolutional layers. Eventually, prior to the fixed last convolution operator, the number of convolutional layers are determined as 2 and their filter numbers are 16 per layer, whose performances are shown in Table 4.18.

This modification visibly enhances the performance of VGG-19 on sagittal radiographs and axial MRIs. Instead of flattening, the last global average pooling calculates the mean

| Modality | Perspective | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| Radiograph | Coronal | 0.8820 | 0.88 | 0.88 | 0.88 | 0.0570 |
| Radiograph | Sagittal | 0.7704 | 0.77 | 0.77 | 0.76 | 0.0894 |
| MRI | Sagittal | 0.6773 | 0.68 | 0.66 | 0.66 | 0.1113 |
| MRI | Axial | 0.6536 | 0.61 | 0.63 | 0.61 | 0.1220 |

**Table 4.18:** Fully convolutional VGG-19 classification and regression performance

| Network | Batch size | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| ResNet-18 | 8 | 0.8380 | 0.83 | 0.83 | 0.83 | 1.5269 |
| ResNet-34 | 16 | 0.8395 | 0.83 | 0.83 | 0.83 | 0.7303 |
| ResNet-50 | 16 | 0.8295 | 0.82 | 0.82 | 0.82 | 3.8338 |
| ResNet-101 | 64 | 0.8220 | 0.82 | 0.81 | 0.80 | 6.9157 |

**Table 4.19:** ResNet classification and regression performance with optimal batch sizes on coronal radiographs

of classes from each pixel. A pixel-wise classification/regression definitely is far more difficult to overfit and reinforces the precision, robustness and reliability of classifiers. That is the reason why with the image size expansion, the upgrade effect is magnifying. The additional convolutional layers slow down the learning rate, in particular nearby the accuracy peak, so that there are more opportunities to fine-tune an ideal model. The less filters are assigned, the slower the convergence is. Hence, the tradeoff between training duration and possible progress settles the filter number as 16. After all, few filters may stagnate the training for redundant epochs to exceed the patience of EarlyStopping.

While the comparison with pre-experiments, the necessity of class distribution balance is empirically proven by the soaring accuracies (around 10%). Moreover, on the balanced datasets, macro average precision, recall and F1-score have approached to their upper limits, since model trainings are not struck into certain class prediction. Naturally, it sets a shining example for the following evaluation.

### 4.4.2 ResNets

Although Keras has pre-trained ResNets, in order to control weight initialization, all the ResNet architectures are trained from sketch with validation accuracy monitoring. The patience of EarlyStopping function is designated to 50, learned from the surge intervals of manifold tryouts. As for batch size, it is tuned by grid search algorithm among 8, 16, 32, 64 and 128 as hyperparameter optimization in aforementioned trails as well, whose results are presented in Table 4.19, Table 4.20, Table 4.22 and Table 4.21 respectively. Although it is tough to discover the rule of optimal batch size, those different batch sizes just give rise to approximately within 1% difference among validation accuracies so that their classification performances with ideal batch size are still comparable. As for regression analysis, our MSEs are calculated per batch by evaluate function in Keras so that their comparability is only applicable to the trainings with the same batch sizes.

On the whole, ResNet-34 accomplishes most outstandingly, which profits from its modest network. This is also in accordance with the discovery of [8]. On the most monotonous dataset, radiographs from coronal plane, smaller structures likewise ResNet-18, have almost the same performance, whereas for images from sagittal view, ResNet-101 occupies

| Network | Batch size | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| ResNet-18 | 64 | 0.6684 | 0.65 | 0.64 | 0.63 | 2.6063 |
| ResNet-34 | 8 | 0.6921 | 0.68 | 0.68 | 0.66 | 3.4473 |
| ResNet-50 | 16 | 0.6691 | 0.64 | 0.65 | 0.63 | 4.4126 |
| ResNet-101 | 8 | 0.6765 | 0.68 | 0.66 | 0.66 | 9.0198 |

**Table 4.20:** ResNet classification and regression performance with optimal batch sizes on sagittal radiographs

| Network | Batch size | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| ResNet-18 | 32 | 0.6554 | 0.64 | 0.62 | 0.59 | 2.8907 |
| ResNet-34 | 16 | 0.6667 | 0.66 | 0.64 | 0.63 | 5.2924 |
| ResNet-50 | 16 | 0.6483 | 0.64 | 0.62 | 0.60 | 5.6758 |
| ResNet-101 | 32 | 0.6483 | 0.64 | 0.63 | 0.63 | 9.7233 |

**Table 4.21:** ResNet classification and regression performance with optimal batch sizes on sagittal MRIs

the second rank, since the shapes of knee joints from sagittal perspective vary in a widen range. Similar with VGGs, ResNet-18, ResNet-34 and ResNet-50 behave almost the same, aiming at axial MRIs. Evidently, due to the structured residual blocks, the accumulation of layers cannot promote feature maps, in particular extracted from axial view. It even may be only connected with shortcuts. Certainly, axial view indeed is not adept in reflecting to JSN.

### 4.4.3 DenseNets

Different with ResNets, DenseNets lift up validation accuracies quite slowly, since its multiple connections within dense blocks. Thereupon, after 300 epochs without any gain for validation accuracy, our EarlyStopping callbacks would restore the optimally trained

| Network | Batch size | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| ResNet-18 | 64 | 0.6165 | 0.55 | 0.58 | 0.53 | 2.8495 |
| ResNet-34 | 16 | 0.6129 | 0.58 | 0.59 | 0.58 | 5.1153 |
| ResNet-50 | 16 | 0.6168 | 0.57 | 0.59 | 0.56 | 5.6195 |
| ResNet-101 | 16 | 0.6053 | 0.52 | 0.57 | 0.51 | 9.9949 |

**Table 4.22:** ResNet classification and regression performance with optimal batch sizes on axial MRIs

| Network | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-121 | 0.8941 | 0.89 | 0.89 | 0.89 | 0.0532 |
| DenseNet-161 | 0.8965 | 0.89 | 0.89 | 0.89 | 0.0505 |
| DenseNet-169 | 0.8920 | 0.89 | 0.89 | 0.89 | 0.0561 |
| DenseNet-201 | 0.8947 | 0.90 | 0.89 | 0.89 | 0.0561 |

**Table 4.23:** DenseNet classification and regression performance on coronal radiographs

| Network | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-121 | 0.7583 | 0.75 | 0.75 | 0.75 | 0.1062 |
| DenseNet-161 | 0.7645 | 0.76 | 0.75 | 0.75 | 0.1083 |
| DenseNet-169 | 0.7583 | 0.75 | 0.75 | 0.75 | 0.1082 |
| DenseNet-201 | 0.7723 | 0.76 | 0.76 | 0.75 | 0.1042 |

**Table 4.24:** DenseNet classification and regression performance on sagittal radiographs

model. In the meanwhile, due to its complex structure, batch size does not have catastrophic impact on the performance, which can be made up by more epochs. For example, DenseNet-201 with 32 samples per batch completes training 1100 epochs or so, while when batch size cuts down to 16, the whole training calls for around 1450 epochs on average. Certainly, among all the contrast tests, the accuracy differences are roughly within 2%. Hence, for the convenience of comparison with VGGs, batch size is arranged to 32.

As the key advantage of DenseNet, the narrow structure with 12 channel growth per layer has been substantiated its marvellous performance in [101], for which our DenseNets keep growth rate as 12. Inspired by skills how to search for ideal filter numbers of VGG convolutional layers, the initial filter number settings are traversed among small integers: 4, 8, 12 and 16, where 8 triumphs over in both accuracy and speed. In general, filter numbers are assigned small enough so that reduction rate is left as the same with the original design (0.5) and Dropout rate is also preserved as 0. In addition to theoretical analysis, there are two more practical arguments for giving up Dropout. Firstly, DenseNets are already trained tardily with direct connections. If bringing in Dropout, the whole training process would be extended out of expectation. Besides, since MOST does not supply as massive images as normal picture databases, even though each group of medical images are expanded by data augmentation, it is still far away from overfitting for such deep networks rather than VGGs, for which compared to DenseNets without Dropout, there is around 5% accuracy less for DenseNets with 0.5 or 0.8 Dropout rate during our hyperparameter optimization trails.

On the radiograph datasets, by virtue of clear image composition, DenseNets act alike and any aforementioned DenseNet can handle with X-Ray images much better than up-to-date researches, like [35] also on the balanced datasets. After all, the connection struc-

| Network | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-121 | 0.6667 | 0.68 | 0.66 | 0.65 | 0.1312 |
| DenseNet-161 | 0.6704 | 0.67 | 0.64 | 0.62 | 0.1275 |
| DenseNet-169 | 0.6667 | 0.69 | 0.66 | 0.65 | 0.1262 |
| DenseNet-201 | 0.6879 | 0.67 | 0.67 | 0.66 | 0.1444 |

**Table 4.25:** DenseNet classification and regression performance on sagittal MRIs

| Network | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-121 | 0.6415 | 0.60 | 0.61 | 0.58 | 0.1323 |
| DenseNet-161 | 0.6584 | 0.63 | 0.63 | 0.61 | 0.1284 |
| DenseNet-169 | 0.6515 | 0.62 | 0.63 | 0.62 | 0.1307 |
| DenseNet-201 | 0.6467 | 0.61 | 0.62 | 0.59 | 0.1637 |

**Table 4.26:** DenseNet classification and regression performance on axial MRIs

ture prevents DenseNets from overfitting. Even if their training accuracies reach 99%, their validation accuracies still maintain at an advanced level. Thus, when training accuracies exceed 99%, a series of attempts are tried out and the optimal one is restored, for which divergent networks perform similarly in the end but with pretty distinctive epoch numbers. This also points out why the patience of EarlyStopping should be set to much higher than other two architectures. If the patience is growing, their performances would continue to be mildly lifted up. Balanced the performance gain and the training duration, 300 epochs is decided. While DenseNet-161 and DenseNet-201 are slightly surpassing respectively on the radiographs, DenseNet-161 and DenseNet-169 are adept in MRIs. Therefore, in general DenseNet-161 can cope with both modality perfectly, in particular in regression analyses. By contrast, DenseNet-201 often obtains the highest accuracy but with also high MSE, as Table 4.25. Apparently, its misclassification offsets more from targets than other networks, for which the clinical diagnosis takes higher risks, although it is more likely that DenseNet-201 offers a correct prediction.

### 4.4.4 FRCNN

As end-to-end learning, hyper-parameters of classifiers such as steps per epoch and initial weights are shared with RPNs depicted in section 4.3.2. Thus, distinguished with other classification algorithms, classifiers of FRCNN are loaded weights from Keras pretrained ResNet-50. Certainly, now that MRIs do not really need RPNs, their bounding box targets are specified to the entire images. Taking the performance of both RPNs and classifiers into consideration, the training process is monitored by the sum of bounding box prediction loss and classification loss. According to the gained knowledge from

| Modality | Perspective | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Radiograph | Coronal | 0.6772 | 0.71 | 0.68 | 0.67 |
| Radiograph | Sagittal | 0.4998 | 0.48 | 0.50 | 0.48 |
| MRI | Sagittal | 0.5620 | 0.48 | 0.56 | 0.49 |
| MRI | Axial | 0.5345 | 0.54 | 0.53 | 0.51 |

**Table 4.27:** FRCNN classification performance

previous experiment observation, if the total loss does not descend within 10 epochs or models are trained beyond 200 epochs, the learning procedure would turn to the end. Inspired by snapshot concept, after each epoch, weights of the trained model is saved, among which the best models (listed in Table 4.27) are opted.

Apparently, ROI proposal has enormous influence on final grading, in particular for X-Ray images. Compared to the performance of its backbone, ResNet-50, no matter from which perspective of radiographs, their accuracies sharply drop over 15%, while those of MRIs slides 8.43% on the average. In order to extract ROIs, shared convolutional layers pay more attention to features describing knee joint contours and then ignore features depicting minor width changes of joint margins which are really beneficial to classifiers. Therefore, MRIs without actual ROI proposal are less impacted.

FRCNN is prevalent for object detection, especially in scenarios where classification can be fulfilled via object outlines. In the field of OA diagnosis, to a great extent, fine distinctions inside subjects are the key to judgement, which is challenging for feature selection of state-of-art FRCNN. In other words, even if the backbone is changed, the outcomes still cannot match with its separated classifier caused by the architecture design.

### 4.4.5 Selection of neural network architectures

Having witnessed the dilemma of end-to-end learning, this section merely sorts out the optimal networks from traditional CNN-based approaches. ResNet is firstly rejected from candidate list. Even though on the MRI datasets, the leading structure, ResNet-34, narrows the classification performance gap between itself and other two architectures to around 1%, it still cannot catch up with their least cases. With respect to radiographs, the disparity between ResNets and DenseNets even comes to 8%. And this difference would be enlarged with data imbalancing as Table 4.28 and Table 4.29. Besides, in regression analyses, compared to VGG architectures and DenseNets, their MSEs are almost one order of magnitude larger.

Generally, the whole performance of VGG-19 is inferior to DenseNets around 1.5% and their gap is increasing with the growth of image size. Revised by replacing fully connected layers, the results of VGG-19 approach to, even surpass, the best DenseNet, now that they work with the similar principle. Hence, served VGG-19 as baseline, DenseNet-

| Architecture | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| VGG-19 | 0.8176 | 0.76 | 0.70 | 0.73 | 0.0779 |
| Revised VGG-19 | 0.8313 | 0.78 | 0.71 | 0.74 | 0.0691 |
| DenseNet-161 | 0.8427 | 0.77 | 0.78 | 0.76 | 0.0709 |
| ResNet-34 | 0.7710 | 0.66 | 0.61 | 0.63 | 2.5173 |

**Table 4.28:** Classification and regression performance on coronal radiographs under JSN scoring system

| Architecture | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| CNN in [28] | 0.634 | 0.62 | 0.61 | 0.58 | —— |
| VGG-19 | 0.6712 | 0.64 | 0.59 | 0.60 | 0.0924 |
| Revised VGG-19 | 0.6626 | 0.63 | 0.60 | 0.60 | 0.0969 |
| DenseNet-161 | 0.6956 | 0.66 | 0.63 | 0.62 | 0.1135 |
| ResNet-34 | 0.6010 | 0.54 | 0.49 | 0.50 | 3.3012 |

**Table 4.29:** Classification and regression performance on coronal radiographs under KL grading system

161 is picked for further comparison with upgraded VGG-19 on the original image sets under both JSN and KL grading systems.

The same with pre-experiments, the imbalance has barely negative effect on the final results. The classification/regression performance differences between JSN and KL are distinguished among coronal radiographs, not only because of one more class from KL, but also owing to the indiscernible grade 1 under KL as previous papers declared [36, 41]. However, no matter under which grading system, DenseNet-161 possesses the overwhelming superiority roundly in classifications, even if compared with up-to-date researches, such as 66.71% accuracy of 5-class classification in [8] and 0.68 precision of 4-class in [36]. The detailed classification reports for each class are listed in Table 4.30-Table 4.32, which exposes the issue for KL grade 1 vividly. On the other hand, VGG architectures own lower MSEs in regression analyses on the whole, although, for the sake of image sizes, revised VGG-19 does not promote so much progress.

On the contrary, revised VGG-19 lifts up both classification and regression performance on X-Ray images from sagittal plane remarkably, even surpasses DenseNet-161 readily under KL assessment system. Visibly, averaging 355 × 568 pixels render a strong robustness of precision, which makes up the vacancy of fully connected layers. Nevertheless, the imbalance starts to make sense on the classification/regression, especially for VGGs, which narrows the performance gap between JSN and KL by weakening grade 1 classification under JSN scoring system as well. After all, from sagittal view, it is more challenging to extract exact features telling apart grade 1 and grade 0.

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.89 | 0.96 | 0.92 |
| 1 | 0.86 | 0.81 | 0.84 |
| 2 | 0.85 | 0.84 | 0.85 |
| 3 | 0.97 | 0.96 | 0.96 |
| Mean | 0.89 | 0.89 | 0.89 |

**Table 4.30:** Class-wise DenseNet-161 classification performance on the balanced coronal radiographs under JSN scoring system

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.89 | 0.95 | 0.92 |
| 1 | 0.62 | 0.39 | 0.48 |
| 2 | 0.80 | 0.83 | 0.81 |
| 3 | 0.76 | 0.94 | 0.84 |
| Mean | 0.77 | 0.78 | 0.76 |

**Table 4.31:** Class-wise DenseNet-161 classification performance on the imbalanced coronal radiographs under JSN scoring system

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.70 | 0.93 | 0.80 |
| 1 | 0.42 | 0.08 | 0.14 |
| 2 | 0.55 | 0.52 | 0.54 |
| 3 | 0.78 | 0.81 | 0.80 |
| 4 | 0.86 | 0.79 | 0.82 |
| Mean | 0.66 | 0.63 | 0.62 |

**Table 4.32:** Class-wise DenseNet-161 classification performance on the imbalanced coronal radiographs under KL grading system

| Architecture | Metric | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| VGG-19 | JSN | 0.6908 | 0.47 | 0.43 | 0.42 | 0.1117 |
| Revised VGG-19 | JSN | 0.7061 | 0.60 | 0.42 | 0.49 | 0.1059 |
| DenseNet-161 | JSN | 0.7389 | 0.62 | 0.49 | 0.51 | 0.1233 |
| VGG-19 | KL | 0.5544 | 0.46 | 0.42 | 0.40 | 0.1192 |
| Revised VGG-19 | KL | 0.5838 | 0.63 | 0.47 | 0.45 | 0.1202 |
| DenseNet-161 | KL | 0.5824 | 0.50 | 0.45 | 0.43 | 0.1481 |

**Table 4.33:** Classification and regression performance on sagittal radiographs

| Architecture | Metric | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| VGG-19 | JSN | 0.6662 | 0.40 | 0.28 | 0.34 | 0.1244 |
| Revised VGG-19 | JSN | 0.6814 | 0.28 | 0.31 | 0.29 | 0.1231 |
| DenseNet-161 | JSN | 0.6830 | 0.53 | 0.35 | 0.35 | 0.1566 |
| VGG-19 | KL | 0.4774 | 0.38 | 0.37 | 0.36 | 0.1382 |
| Revised VGG-19 | KL | 0.5176 | 0.27 | 0.31 | 0.28 | 0.1282 |
| DenseNet-161 | KL | 0.5097 | 0.40 | 0.33 | 0.36 | 0.1866 |

**Table 4.34:** Classification and regression performance on sagittal MRIs

| Architecture | Metric | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|---|
| VGG-19 | JSN | 0.6603 | 0.58 | 0.31 | 0.40 | 0.1232 |
| Revised VGG-19 | JSN | 0.6702 | 0.65 | 0.31 | 0.42 | 0.1196 |
| DenseNet-161 | JSN | 0.6763 | 0.39 | 0.39 | 0.39 | 0.1652 |
| VGG-19 | KL | 0.4943 | 0.38 | 0.29 | 0.33 | 0.1321 |
| Revised VGG-19 | KL | 0.5172 | 0.39 | 0.32 | 0.35 | 0.1262 |
| DenseNet-161 | KL | 0.5215 | 0.43 | 0.36 | 0.39 | 0.1728 |

**Table 4.35:** Classification and regression performance on axial MRIs

Feature selection troubles MRIs from sagittal perspective more seriously, even involving grade 2 of KL, since they display more knee details indirectly related to JSN. In addition, how to construct features expressing few samples of grade 3 under JSN or grade 4 under KL scoring system is highly demanding. Thus, recalls fall off sharply and the gap between performances of different grading systems is sealed. With the smallest image shape, convolution replacement clearly cannot boost the performance of VGG-19, even becomes a burden on classification. Although the accuracies are enhanced, it relies on grade 0 and grade 2 (JSN)/grade 3 (KL). Then the unsatisfied macro means can be easily imaged. The outputs of DenseNet-161 also tend to grade 0 and grade 2 (JSN)/grade 3 (KL) but their outcomes are supported by the class with fewest images, which retards the performance decline. Thus, in practical, contributed by dense blocks, overcoming the lack of data is easier for classifiers than extracting features directing at tiny lesions.

This also clarifies why axial MRIs can be diagnose as grade 3 (JSN)/grade 4 (KL), however, grade 1 and grade 2 cannot be recognized roughly. Pixel-wise classification/regression also conduces to the identification of class with fewest data by VGG-19 but accurate features from axial plane for joint margin portrayal are scarce. Hence, although the overall MRI performances under both semi-quantitative assessment criteria are similar, their inner obstacles for both perspectives are widely divergent.

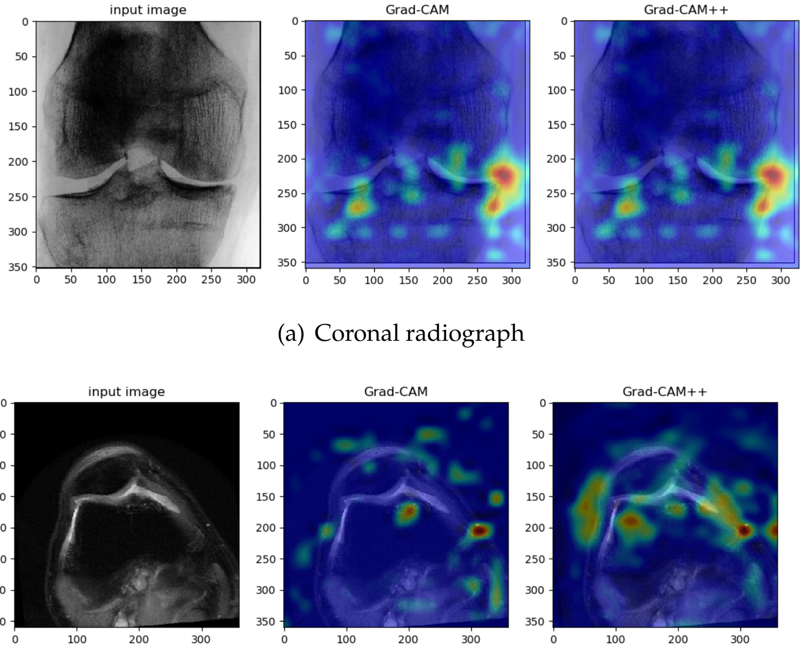
DenseNet-161 architecture has stronger resistance to imbalanced image sets so that its classification performance properly maintains at the highest level, compared with other architectures. However, with the drop of accuracy, its MSEs raise more rapidly than revised VGG-19, even VGG-19. Distinctly, the misclassification of DenseNets is easier with large offsets, especially for MRIs. In addition, with the expansion of image sizes, the modification of VGG-19 plays a more leading role in classification, while the memory limitation to DenseNet comes into play, for which the ideal choice turns into revised VGG-19.

## 4.5 Decision visualization

Precise decisive feature localization is vital not only for model prediction explanation but also for rapidly confirming the reliability of the outcome, especially for potentially false positive cases [77]. Therefore, a more trustworthy attention map calculation approach is selected between Grad-CAM and Grad-CAM++ by our manual judgement. Due to external conditions, this thesis fails to verify the localization accuracy with radiologists.

Primarily, the more accurate the model is, the more alike the visualizations between Grad-CAM and Grad-CAM++, since key features are then more simply identified. Hence, their visualizations of coronal radiographs are the same for correct predictions, even wrong grading, whereas visualizations for the same axial MRI slice vary in a board range, even if they are classified exactly, as fig. 4.4. In order to spotlight entire objects, instead of their parts, Grad-CAM++ replaces global averaging by the weighted mean, for which in our case, Grad-CAM++ highlights more conjoined features precisely (fig. 4.4(b) and fig. 4.5(b)). As previous papers focus on radiographs from coronal plane, Grad-CAM is enough for attention map generation. On account of assisting diagnosis based on multimodality, Grad-CAM++ is adopted to our pipeline, by virtue of its extensive applicability.

As shown in fig. 4.5, the saliency maps not only verify the reliability of our models, but also serve as a valuable tool for issue exposure. For example, the misclassified X-Ray images from coronal plane in common highlight the upper or bottom borders e.g. the heatmap generated by the Grad-CAM as shown in fig. 4.5(b). Even if the classification outputs are aligned with the targets, the borders are also spotlighted as shown in fig. 4.5(d). Unacquainted with the knee joint structure, our ROI annotation covers extra shaft part of femurs or tibiae, especially femurs are cut out in terms of the shadow of patella which is not directly related to OA quantification. Thereupon, non-uniform lengths of bones increase the possibility of misclassification, which is however, caused by human factor rather than the whole pipeline itself. Consequently, ROI is merely delineated in the condyle region, which is contributed by saliency maps.



(a) Coronal radiograph

(b) Axial MRI

**Figure 4.4:** The visualization comparison between Grad-CAM and Grad-CAM++

(a) Target JSN: Grade 2, Prediction: Grade 0; (b) Target JSN: Grade 3, Prediction: Grade 3

## 4.6 Model ensemble

No matter which networks are employed, the definite trend of classification levels between perspectives per modality is stable. Radiographs from coronal plane possesses impressive classification capability, leading at least 14% accuracy together with other evaluation metrics. Certainly, less features in the images contributes a lot to its success. However, the key to its triumph is the labels stemming from scoring systems, especially JSN, which lays more emphasis on the width of knee joint margin. Compared to other perspectives, coronal view can exhibit joint margins most clearly, which explain why previous papers have special preference on that. Hence, our model ensemble also strengthen the usage of images from this group.

Ranking at the second place, sagittal X-Ray images benefit by minor interference by extra features. MRIs encompass a great deal of OA features, however, they are merely involved into more sophisticated semi-quantitative or quantitative assessments. This is the reason why MRIs cannot exert its advantages under KL or JSN scoring systems. Then, obviously, MRIs from axial perspective which cannot give expression to the width of knee joint margin but hold considerable additional OA features serve as the last option for our

| Architecture | Accuracy | MSE |
|---|---|---|
| DenseNet-161 | 0.8965 | 0.1210 |
| DenseNet-169 | 0.8920 | 0.1174 |
| DenseNet-201 | 0.8947 | 0.1071 |
| VGG19 | 0.8896 | 0.1228 |

**Table 4.36:** Classification and regression performance for top models on coronal radiographs

| Architecture combination | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-161+DenseNet-201 | 0.8956 | 0.89 | 0.89 | 0.89 | 0.1095 |
| DenseNet-161+DenseNet-169 | 0.8896 | 0.89 | 0.88 | 0.88 | 0.1138 |
| DenseNet-161+VGG19 | 0.8908 | 0.89 | 0.89 | 0.89 | 0.1110 |
| DenseNet-161+DenseNet-201+ DenseNet-169 | 0.8890 | 0.89 | 0.88 | 0.88 | 0.1162 |
| DenseNet-161+DenseNet-201+ VGG19 | 0.8844 | 0.88 | 0.88 | 0.88 | 0.1216 |
| DenseNet-161+DenseNet-169+ VGG19 | 0.8832 | 0.88 | 0.88 | 0.88 | 0.1210 |
| DenseNet-161+DenseNet-201+ DenseNet-169+VGG19 | 0.8784 | 0.87 | 0.88 | 0.87 | 0.1276 |
| DenseNet-201+DenseNet-169 | 0.8947 | 0.89 | 0.89 | 0.89 | 0.1113 |

**Table 4.37:** Classification and regression performance for model ensemble by prediction maximization on coronal radiographs

multimodality integration.

### 4.6.1 Model ensemble with same modality

Following above analysis, model ensemble is carried on models classifying radiographs from coronal view. Top 3 models based on the whole performance (directly calculated on the entire validation datasets, rather than batch by batch, as shown in Table 4.36) are picked: DenseNet-161, DenseNet-169 and DenseNet-201. In consideration of the variety of architectures and regression analysis conclusion, VGG-19 is also involved into model ensemble, whose outcomes are presented in Table 4.37 and Table 4.38.

No matter how to combine architectures, assembling by averaging their Softmax posteriors indeed promotes the classification and regression performance roundly. Taking the maximum among predictions as final decision is easily influenced by outliers with high

| Architecture combination | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-161+DenseNet-201 | 0.8992 | 0.90 | 0.89 | 0.89 | 0.1059 |
| DenseNet-161+DenseNet-169 | 0.8989 | 0.90 | 0.89 | 0.89 | 0.1071 |
| DenseNet-161+VGG19 | 0.9059 | 0.90 | 0.90 | 0.90 | 0.1026 |
| DenseNet-161+DenseNet-201+ DenseNet-169 | 0.9040 | 0.90 | 0.90 | 0.90 | 0.0993 |
| DenseNet-161+DenseNet-201+ VGG19 | 0.9122 | 0.91 | 0.91 | 0.91 | 0.0878 |
| DenseNet-161+DenseNet-169+ VGG19 | 0.9113 | 0.91 | 0.91 | 0.91 | 0.0929 |
| DenseNet-161+DenseNet-201+ DenseNet-169+VGG19 | 0.9077 | 0.90 | 0.90 | 0.90 | 0.0957 |
| DenseNet-201+DenseNet-169 | 0.8959 | 0.89 | 0.89 | 0.89 | 0.1068 |

**Table 4.38:** Classification and regression performance for model ensemble by Softmax posterior average on coronal radiographs

scores. To a great extent, the mean possibilities for each class ameliorates the affect of outliers so that this thesis adopts Softmax posterior average as ensemble methodology.

On the whole, VGG-19 does bring in substantial improvement, which turns out that the architecture mixture assists to breakthrough and the original performance itself is not the only key to eventual outputs. After all, models in the same series behave similarly so that their extracted features as well as predictions are alike. Accordingly, aiming at those samples difficult to classify, their outcomes cannot make up each other, which presents the significance of model diversity. In addition, the notion that the more models are assembled, the better outcomes are attained is dispelled by our empirical results.

The combination of DenseNet-161, DenseNet-201 and VGG-19 visibly precedes other groups, even it is associated with less basic models. Analysing its confusion matrix (fig. 4.6) in depth, the misclassified samples are gathered on the adjacent classes of ground truths, for which its MSE is as minor as 0.0878, exactly conforming to the clinical expectation. Compared Table 4.39 with Table 4.30, there is no doubt that the correction effect on grade 1 and grade 2 is notable. Those misclassified grade 2 X-Ray images are adjusted back to grade 1. In reality, one more grade prediction is acceptable. In this case, the recall for each class would be dramatically elevated to 1.00, 0.90, 0.87 and 0.98, which in principle achieves the practical use requirements.

When it comes to the imbalanced dataset under JSN scoring, the precision, recall and F1-score of the ensemble of DenseNet-161 and revised VGG-19 are 0.80, 0.79 and 0.77 respectively. Clearly, thanks to the adequate space for lifting up, even if employing the prediction maximization method, each evaluation metric is still boosted 2% and surpasses

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.90 | 0.96 | 0.93 |
| 1 | 0.86 | 0.85 | 0.86 |
| 2 | 0.90 | 0.83 | 0.87 |
| 3 | 0.98 | 0.98 | 0.98 |
| Mean | 0.91 | 0.91 | 0.91 |

**Table 4.39:** Class-wise classification performance of the model ensemble on the balanced coronal radiographs

| Modality combination | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| Coronal radiograph | 0.8799 | 0.88 | 0.88 | 0.88 | 0.1394 |
| Coronal radiograph+Sagittal radiograph | 0.8848 | 0.89 | 0.89 | 0.89 | 0.1317 |
| Coronal radiograph+Sagittal MRI | 0.8813 | 0.89 | 0.88 | 0.88 | 0.1461 |
| Coronal radiograph+Axial MRI | 0.8858 | 0.89 | 0.89 | 0.89 | 0.1272 |

**Table 4.40:** Classification and regression performance for multimodality integration (2 image sets)

up-to-date researches readily. All in all, facilitated by model ensemble, our pipeline has fulfilled the need of clinical applications by overall 91% grading performance from coronal perspective.

### 4.6.2 Multimodality integration

As analyzed in the beginning of this section, models based on radiographs from coronal plane are the cornerstone of our multimodality integration so that the optimal trained model, DenseNet-161, is fixed for comparison and the top models trained from other 3 image sets are fused into by Softmax posterior average, whose effect has been corroborated by section 4.6.1. Inspired by the conclusion of model ensemble, models from other perspectives are picked from VGG series to reinforce the entire model variety. Since the random patient selection of grade 3 for dataset balance draws in different subjects and then there are merely 462 grade 3 images in common, the performance of DenseNet-161 is recalculated in Table 4.40.

Noticeably, the statistical slump in comparison to the balanced dataset results from the data deficiency of grade 3. Even if the accuracy differences between models trained from coronal X-Ray images and other perspectives are considerable, multimodality integration still contributes to prediction enhancement with the imbalanced image distribution,

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.88 | 0.96 | 0.92 |
| 1 | 0.86 | 0.81 | 0.84 |
| 2 | 0.85 | 0.84 | 0.85 |
| 3 | 0.94 | 0.92 | 0.93 |
| Mean | 0.88 | 0.88 | 0.88 |

**Table 4.41:** The baseline of class-wise multimodality integration classification performance under JSN scoring system

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.88 | 0.97 | 0.92 |
| 1 | 0.88 | 0.80 | 0.84 |
| 2 | 0.85 | 0.85 | 0.85 |
| 3 | 0.94 | 0.92 | 0.93 |
| Mean | 0.89 | 0.89 | 0.89 |

**Table 4.42:** Class-wise multimodality integration classification performance based on coronal radiographs and sagittal radiographs under JSN scoring system

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.86 | 0.98 | 0.92 |
| 1 | 0.87 | 0.80 | 0.84 |
| 2 | 0.87 | 0.83 | 0.85 |
| 3 | 0.95 | 0.92 | 0.94 |
| Mean | 0.89 | 0.88 | 0.88 |

**Table 4.43:** Class-wise multimodality integration classification performance based on coronal radiographs and sagittal MRIs under JSN scoring system

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.88 | 0.98 | 0.93 |
| 1 | 0.89 | 0.80 | 0.84 |
| 2 | 0.86 | 0.85 | 0.85 |
| 3 | 0.94 | 0.93 | 0.93 |
| Mean | 0.89 | 0.89 | 0.89 |

**Table 4.44:** Class-wise multimodality integration classification performance based on coronal radiographs and axial MRIs under JSN scoring system

| Additional models | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| DenseNet-169 | 0.8925 | 0.90 | 0.90 | 0.90 | 0.1113 |
| SR(D) | 0.8918 | 0.90 | 0.90 | 0.90 | 0.1124 |
| DenseNet-169+SR(D) | 0.8936 | 0.90 | 0.90 | 0.90 | 0.1064 |
| SR(V) | 0.8946 | 0.90 | 0.90 | 0.90 | 0.1092 |
| DenseNet-169+SR(V) | 0.8946 | 0.90 | 0.90 | 0.90 | 0.1054 |
| SM(D) | 0.8922 | 0.90 | 0.90 | 0.90 | 0.1110 |
| DenseNet-169+SM(D) | 0.8950 | 0.90 | 0.90 | 0.90 | 0.1071 |
| SM(V) | 0.8957 | 0.90 | 0.90 | 0.90 | 0.1071 |
| DenseNet-169+SM(V) | 0.8967 | 0.90 | 0.90 | 0.90 | 0.1033 |
| AM(D) | 0.8929 | 0.90 | 0.90 | 0.90 | 0.1120 |
| DenseNet-169+AM(D) | 0.8932 | 0.90 | 0.90 | 0.90 | 0.1078 |
| AM(V) | 0.8939 | 0.90 | 0.90 | 0.90 | 0.1071 |
| DenseNet-169+AM(V) | 0.8950 | 0.90 | 0.90 | 0.90 | 0.1061 |

**Table 4.45:** Classification and regression performance for multimodality integration based on model ensemble outcome (2 image sets)

D: DenseNet, V: VGG, SR: Sagittal radiograph, SM: Sagittal MRI, AM: Axial MRI, which is the same with Table 4.46 and Table 4.47.

especially for grade 0. Seen in Table 4.41-4.44, the recalls of grade 0 are even higher than the selected assembled model in section 4.6.1 and precisions of other 3 grades also lift up slightly. Compared with model based on MRIs from sagittal plane, other two models have a more appreciable effect, which may benefit from revised VGG-19.

By means of the wider architecture discrepancy, revised VGG-19 remedies the drop of performance among the three models. Concerning the information entropy provided by image itself, axial view with totally distinctive features prevails undoubtedly. In order to further upgrade multimodality integration, the key to the success of model ensemble, DenseNet-161, DenseNet-201 and VGG-19 (Base), are taken advantage of. Top models of other 3 image sets obtained from both DenseNets and VGGs are melded into.

Evidently, multimodality integration do assists to prediction optimization on the basis of our model ensemble (Table 4.45). Although the differences among modalities are insignificant, overall performance integrated with models from sagittal MRIs is the optimal, followed by models acquired from sagittal X-Ray images. Clearly, rather than single model combinations, model diversification supported by models attained from coronal radiographs has filled the achievement gap between VGG-19 and revised VGG-19, for which the vital factor roots in the original accuracy and the image information entropy.

Thus, consolidating basic accuracy through adding one more model based on X-Ray im-

| Additional models | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| SR(D)+SM(D) | 0.8932 | 0.90 | 0.90 | 0.90 | 0.1089 |
| SR(D)+AM(D) | 0.8929 | 0.90 | 0.90 | 0.90 | 0.1082 |
| SR(D)+SM(V) | 0.8946 | 0.90 | 0.90 | 0.90 | 0.1064 |
| SR(D)+AM(V) | 0.8939 | 0.90 | 0.90 | 0.90 | 0.1071 |
| SR(V)+SM(D) | 0.8950 | 0.90 | 0.90 | 0.90 | 0.1061 |
| SR(V)+AM(D) | 0.8915 | 0.90 | 0.89 | 0.89 | 0.1113 |
| SR(V)+SM(V) | 0.8939 | 0.90 | 0.90 | 0.90 | 0.1089 |
| SR(V)+AM(V) | 0.8939 | 0.90 | 0.90 | 0.90 | 0.1061 |
| SM(D)+AM(D) | 0.8897 | 0.90 | 0.89 | 0.89 | 0.1145 |
| SM(V)+AM(D) | 0.8925 | 0.90 | 0.90 | 0.90 | 0.1075 |
| SM(D)+AM(V) | 0.8936 | 0.90 | 0.90 | 0.90 | 0.1085 |
| SM(V)+AM(V) | 0.8935 | 0.90 | 0.90 | 0.90 | 0.1075 |

**Table 4.46:** Classification and regression performance for multimodality integration based on model ensemble outcome (3 image sets)

ages from coronal plane makes sense under this condition. By contrast, it is redundant for model ensemble of single perspective. In view of Base composition (2 DenseNets and 1 VGG-19), naturally, fusion with top VGGs trained on other 3 datasets yields better outcomes. Comprehensively, Base + DenseNet-169 (coronal radiograph) + VGG-19 (sagittal MRI) thoroughly deserves to be settled on the ideal integration. With regards to assembling models generated on 3 or 4 image sets, the group of DenseNet-161, DenseNet-169, DenseNet-201 and VGG-19 is set as the New Base for further assessment, stemming from their aforementioned accomplishment.

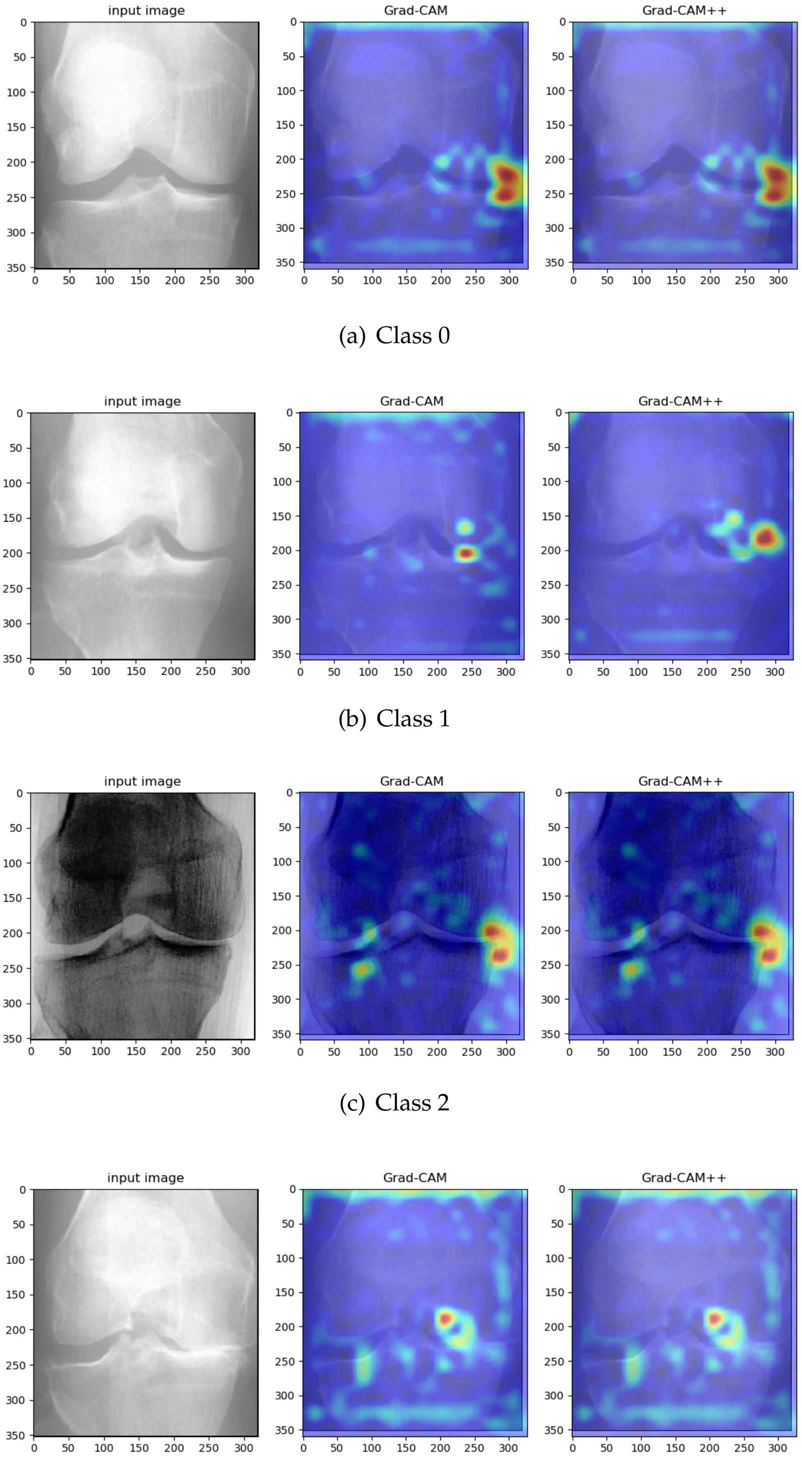
Although fusing more modalities shortens their performance differences, Table 4.46 comes up with solid evidence again that X-Ray images and MRIs from sagittal perspective are more appropriate for multimodality integration established on New Base. Their architecture-mixed combinations, both SR(V) + SM(D) and SR(D) + SM(V), occupy the leading rankings among 3-Modality Integration. Discriminating from integration based on 2 image sets, its promotion effect is relatively meager, in particular for VGG series models. After all, there is scarce space for upsurge.

Due to the comparatively low original accuracies, rather than 3 image sets, integrating models trained from 4 image sets does not benefit to prediction. The performance tendency among various number of model integration has already revealed that models learned on 3 datasets has offered enough information for JSN score classification/regression so that more predicted class possibilities, in particular generated by AM(D), just bring in interferences. In summary, balancing the cost of data acquisition and its possible performance gain, our pipeline singles out the group of DenseNet-161, DenseNet-169, DenseNet-201, VGG-19 and SM(V) as multimodality integration module,

| Additional models | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| SR(D)+SM(D)+AM(D) | 0.8869 | 0.90 | 0.89 | 0.89 | 0.1236 |
| SR(D)+SM(D)+AM(V) | 0.8929 | 0.90 | 0.90 | 0.90 | 0.1145 |
| SR(D)+SM(V)+AM(D) | 0.8915 | 0.90 | 0.89 | 0.90 | 0.1149 |
| SR(D)+SM(V)+AM(V) | 0.8932 | 0.90 | 0.90 | 0.90 | 0.1089 |
| SR(V)+SM(D)+AM(D) | 0.8876 | 0.90 | 0.89 | 0.89 | 0.1205 |
| SR(V)+SM(D)+AM(V) | 0.8939 | 0.90 | 0.90 | 0.90 | 0.1082 |
| SR(V)+SM(V)+AM(D) | 0.8904 | 0.90 | 0.89 | 0.89 | 0.1124 |
| SR(V)+SM(V)+AM(V) | 0.8922 | 0.90 | 0.90 | 0.90 | 0.1089 |

**Table 4.47:** Classification and regression performance for multimodality integration based on model ensemble outcome (4 image sets)

which profits from ample information absorbed from both modalities and perspectives. Certainly, limited by the access to medical images, our scheme is also furnished with model ensemble module operated via averaging Softmax posteriors of DenseNet-161, DenseNet-201 and VGG-19 trained on coronal radiographs.



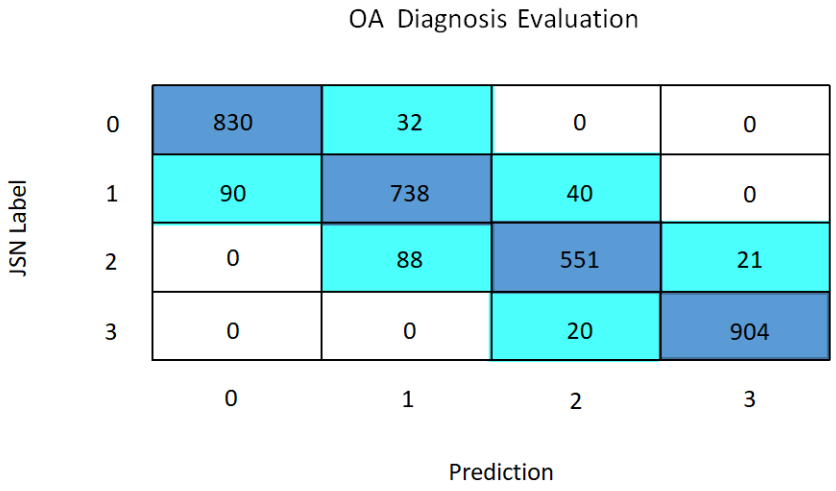
(a) Class 0

(b) Class 1

(c) Class 2

(d) Class 3

**Figure 4.5:** Class-wise decision visualization on the balanced coronal radiographs



OA Diagnosis Evaluation

| JSN Label | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 830 | 32 | 0 | 0 |
| 1 | 90 | 738 | 40 | 0 |
| 2 | 0 | 88 | 551 | 21 |
| 3 | 0 | 0 | 20 | 904 |

Prediction

**Figure 4.6:** Confusion matrix of the assembled model

# Chapter 5

# Conclusion and Outlook

Targeting to advance the state-of-art automatic knee OA diagnosis by upgrading deep learning based architectures, this thesis establishes the pipeline with preprocessing, ROI detection, classification/regression, decision visualization and ensemble modules on the multimodality integration concept, in terms of our investigation on OA quantization criteria and its automatic detection mechanisms. Having drawn a comprehensively theoretical comparison among plentiful medical image algorithms for each module, the selected approaches are evaluated on plain radiographs and MRIs in MOST public database.

Eventually, an innovative framework assembling multi-models trained from coronal radiographs and sagittal MRIs is proposed to break through existing modality restrictions, whose backbone consists of U-Net ROI detection and classification/regression by DenseNets together with VGG-19. In view of supplying reliability and transparency of the grading procedure, class-discriminating attention maps are generated by Grad-CAM++. Compared with up-to-date knee OA quantification researches, our validation under both OARSI JSN and KL scoring system confirms its superior classification accuracy.

## 5.1 Discussion

Having witnessed the growing demand of automatic knee OA quantification, this thesis reviews thoroughly previous studies on computer-aid OA assessments, from which three technical bottlenecks are exposed: noises, artefacts and intrinsic modality limitations. Aiming at above three points to be broken through, the basic workflow is settled on, including preprocessing, ROI detection, classification/regression and multimodality integration module. After inquiring deeply into approaches for each module together with their evaluation metrics, a novel module for decision visualization is brought in to highlight significant features for classification and a comparative feasibility analysis is given to nominate algorithm candidates for our scheme, which is validated on MOST cohort.

Served as pre-experiments, a series of radiographic preprocessing approaches grouped histogram equalization, Perona-Malik filtering and unsharp masking edge enhancement with various parameters are compared by their final classification and regression performance of VGG-19, while the optimal parameter combination of Perona-Malik filtering, unsharp masking and MRI slice integration is sorted out for the MRI preprocessing approach. Ultimately, the bundles, histogram equalization + slight Perona-Malik filtering ($K$=50, iterations=1) and slight Perona-Malik filtering ($K$=50, iterations=1) + unsharp masking with EDGE_ENHANCE Kernel + MRI slice averaging stood out respectively for both modalities. On the basis of enhanced radiographs, ROI detection trails are conducted among FCN, U-Net and RPN, where U-Net built on ResNet-18 outshines others from precision aspect, whereas RPN transcends by its direct coordinate learning. Thanks to the architecture superiority, both methods exceed state-of-art researches.

Within extracted ROIs, classification/regression module draws an all-round comparison under both JSN and KL grading system among FRCNN, VGG, ResNet and DenseNet series, in which DenseNet-161 excels by stably high accuracies on both perspectives of radiograph and MRI. However, with the expansion of image sizes, the structure advantages of our revised VGG-19 are reflected to its regression preeminence, rather than DenseNet-161. In brief, thanks to the dataset balancing and sensitive JSN scoring system, our single model classification on coronal view of X-Ray images achieves 89.65% accuracy, together with 0.89 precision, recall and F1-score, which surpasses current publications at least 15%. Even if on imbalanced image sets, there is at least a 5% performance gain for our DenseNet-161. Furthermore, its discriminative features are spotlighted by Grad-CAM++ exactly, which verifies the reliability of our prediction.

Served as base of multimodality integration, top models (DenseNet-161, DenseNet-201 and VGG-19) trained on the most suitable image set, balanced radiographs from coronal plane, are assembled by Softmax posterior averaging, which accomplishes 91.22% accuracy and 0.0878 MSE, along with 0.91 precision, recall and F1-score. On the imbalanced dataset, its precision, recall and F1-score are still 0.80, 0.79 and 0.77 respectively. When it comes to multimodality integration, the class distribution turns into imbalanced, due to the inconsistency of assessment subjects. However, the combination of model ensemble (DenseNet-161, DenseNet-169, DenseNet-201 and VGG-19) on coronal X-Ray images and VGG-19 attained from sagittal MRIs still boosts the whole performance metrics to 0.90, followed by the model ensemble fusing with revised VGG-19 learned on sagittal radiographs.

The tradeoff between the cost of data acquisition as well as further processing and the possible performance gain determines the group of radiographs from coronal plane and MRIs from sagittal view as the optimal solution, while model ensemble of DenseNet-161, DenseNet-201 and VGG-19 trained by coronal radiographs is also perfect alternative. In the meanwhile, the significance of model variety for model ensemble is verified by the superior results of the balanced prediction mixture of DenseNets and VGGs.

## 5.2 Limitation

Although our proposed approach has already satisfied the clinical demand, there are still potential spaces to enhance its reliability, which can be outlined as follows:

- **MSE standardization:** Our regression analyses for single models are based on the outputs of Keras evaluate function, which calculates MSEs per batch. In this case, although they are still comparable, their average MSEs are more likely to have biases than the direct computation on the entire validation datasets. Therefore, MSEs for single models would be more convictive to be calculated directly on validation sets as what model ensembles do.
- **ROI calibration:** As analysed in section 4.5, due to the lack of medical expertise, parts of misclassification results from extra region demarcation for ROIs. Thus, ROI calibration in terms of the decision visualization outputs or even radiologists' suggestions would be beneficial to refine final performance.
- **Validation based on other databases:** Likewise [8, 28], they trained models on MOST cohort, while their evaluation carried out on OAI database. Apparently, under this occasion, their outcomes possess higher universality and robustness. However, limited by the access to OAI or other data sources, this validation only can be conducted when other resources are available in the future.

## 5.3 Future work

On the basis of our satisfying pipeline, there are several directions to further extend its functionalities in real practice, even more application areas, which are outlined in the following. Although Grad-CAM++ identifies critical features for classification, it does not explain the relationship between grading and those features. [77] framed the concept of prediction basis that present the typical elements for the relative class contained in examined medical images. Distinctly, appending prediction bases to saliency maps generated by Grad-CAM++ would offer end users an better explanation of how our future online diagnosis application makes prediction.

Although KL and JSN scoring system own wide applicability, they cannot make fully use of OA features contained in MRIs, since they are designed for plain radiographs, which ignore a great deal of details. MRIs frequently detect further cartilage loss and fluctuation of bone marrow lesions, effusion, synovitis and Hoffa-synovitis at the follow-up [16]. Thus, KL grade 4 or JSN score 3 can still deteriorate and the term "end stage" seems to be no longer appropriate [16]. If there would be grading systems taking advantage of MRI details and compatible for both X-Ray images and MRIs, the effect of multimodality integration would be maximized on the basis of current stage.

Predictions are more accurate established on patients' age, sex, Body-Mass Index, given knee injury or surgery history, symptomatic assessment report and so on, since their outputs refer to more decisive factors [36]. Certainly, it greatly depends on the collection of meta-data. Weighing the effort of data acquisition and possible profits, time dimension import is an optimal alternative. The diagnosis of certain visit is heavily influenced by its previous grade, for which availing of follow-up visits of the same patient makes sense for classification refinement even under complex assessment systems. In future we intend to consider these scenario as well as additional factors and parameters.

# Appendix A

# Abbreviation

| | |
|---|---|
| Adam | Adaptive moment estimation |
| ANN | Artificial Neural Networks |
| AUC | Area Under the receiver operating characteristic Curves |
| BCE | Binary Cross Entropy |
| BLOKS | Boston Leeds Osteoarthritis Knee Score |
| CAM | Class Activation Maps |
| CNN | Convolutional Neural Network |
| CT | Computed Tomography |
| DenseNet | Dense Convolutional Network |
| DICOM | Digital Imaging and Communications in Medicine |
| EM | Expectation-Maximization |
| FCN | Fully Convolutional Neural Network |
| FN | False Negative |
| FP | False Positive |
| FRCNN | Faster Region Convolutional Neural Network |
| FSA | Fractal Signature Analysis |
| GA | Genetic Algorithm |
| GAN | Generative Adversarial Network |
| GHMM | Gaussian Hidden Markov Model |
| GLCM | Gray-Level Co-Occurrence Matrix |
| GLM | General Linear Model |
| GMM | Gaussian Mixture Model |
| IoU | Intersection-over-Union |
| JSW | Joint Space Width |
| JSN | Joint Space Narrowing |

| | |
|---|---|
| KIMRISS | Knee Inflammation MRI Scoring System |
| KL | Kellgren and Lawrence |
| KNN | K Nearest Neighbor |
| KOSS | Knee Osteoarthritis Scoring System |
| LBP | Local Binary Pattern |
| LDIC | Location-Dependent Image Classification |
| MAE | Mean Absolute Error |
| MF | Minkowski Functionals |
| ML | Machine Learning |
| MOAKS | Osteoarthritis Knee Score |
| MOST | Multicenter Osteoarthrithis Study |
| MRI | Magnetic Resonance Imaging |
| MSE | Mean Squared Error |
| OA | Osteoarthritis |
| OAI | Osteoarthritis Initiative |
| OARSI | Osteoarthritis Research Society International |
| PCA | Principal Component Analysis |
| RBF | Radial Basis Function |
| RBM | Restricted Boltzmann Machine |
| RCNN | Region-CNN |
| ReLU | Rectified Linear Unit |
| ResNet | Residual Network |
| ROI | Region of Interest |
| RPN | Region Proposal Network |
| RW | Random-Walker |
| SGD | Stochastic Gradient Descent |
| SOM | Self Organizing Map |
| STM | Stereological and Textural Measurements |
| SVM | Support Vector Machines |
| SVR | Support Vector Regression |
| TP | True Positive |
| TN | True Negative |
| WORMS | Whole-Organ Magnetic Resonance Imaging Score |
| YLDs | Years Lived with Disability |

# Bibliography

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

[2] Pooja P Kawathekar and Kailash J Karande. Severity analysis of osteoarthritis of knee joint from x-ray images: A literature review. In *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on*, pages 648–652. IEEE, 2014.

[3] Jaume Puig-Junoy and Alba Ruiz Zamora. Socio-economic costs of osteoarthritis: a systematic review of cost-of-illness studies. In *Seminars in arthritis and rheumatism*, volume 44, pages 531–541. Elsevier, 2015.

[4] Selvon F St Clair, Carlos Higuera, Viktor Krebs, Nabil A Tadross, Jerrod Dumpe, and Wael K Barsoum. Hip and knee arthroplasty in the geriatric population. *Clinics in geriatric medicine*, 22(3):515–533, 2006.

[5] Theo Vos, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Z Chen, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, 388(10053):1545–1602, 2016.

[6] Andrew A Guccione, David T Felson, Jennifer J Anderson, John M Anthony, Yuqing Zhang, PW Wilson, Margaret Kelly-Hayes, Philip A Wolf, Bernard E Kreger, and William B Kannel. The effects of specific medical conditions on the functional limitations of elders in the framingham study. *American journal of public health*, 84(3):351–358, 1994.

[7] Sanjay Gupta, GA Hawker, A Laporte, R Croxford, and PC Coyte. The economic burden of disabling hip and knee osteoarthritis (oa) from the perspective of individuals living with this condition. *Rheumatology*, 44(12):1531–1537, 2005.

[8] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.

[9] Yaodong Du, Juan Shan, and Ming Zhang. Knee osteoarthritis prediction on mr images using cartilage damage index and machine learning methods. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 671–677. IEEE, 2017.

[10] Yosuke Uozumi, Kouki Nagamune, and Kiyonori Mizuno. Computer-aided segmentation system of posterior cruciate ligament in knee joint from ct and mri using anatomical information: A pilot study of system configuration. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 2295–2298. IEEE, 2015.

[11] V Ashwin Kumar and AK Jayanthy. Classification of mri images in 2d coronal view and measurement of articular cartilage thickness for early detection of knee osteoarthritis. In *Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on*, pages 1907–1911. IEEE, 2016.

[12] Holger Jahr, Nicolai Brill, and Sven Nebelung. Detecting early stage osteoarthritis by optical coherence tomography? *Biomarkers*, 20(8):590–596, 2015.

[13] Mikko AJ Finnilä, Jérôme Thevenot, Olli-Matti Aho, Virpi Tiitu, Jari Rautiainen, Sami Kauppinen, Miika T Nieminen, Kenneth Pritzker, Maarit Valkealahti, Petri Lehenkari, et al. Association between subchondral bone structure and osteoarthritis histopathological grade. *Journal of Orthopaedic Research*, 35(4):785–792, 2017.

[14] Artjoms Suponenkovs, Zigurds Markovics, and Ardis Platkajis. Knee-joint tissue recognition in magnetic resonance imaging. In *Neumann Colloquium (NC), 2017 IEEE 30th*, pages 000041–000046. IEEE, 2017.

[15] Travis B Smith and Krishna S Nayak. Mri artifacts and correction strategies. *Imaging in Medicine*, 2(4):445–457, 2010.

[16] Alexander Mathiessen, Marco Amedeo Cimmino, Hilde Berner Hammer, Ida Kristin Haugen, Annamaria Iagnocco, and Philip G Conaghan. Imaging of osteoarthritis (oa): What is new? *Best Practice & Research Clinical Rheumatology*, 30(4):653–669, 2016.

[17] JC Buckland-Wright. Quantitative radiography of osteoarthritis. *Annals of the rheumatic diseases*, 53(4):268, 1994.

[18] Ahmad Fadzil Mohd Hani, Aamir Saeed Malik, Dileep Kumar, Raja Kamil, Ruslan Razak, and Azman Kiflie. Features and modalities for assessing early knee osteoarthritis. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–6. IEEE, 2011.

[19] Dileep Kumar, Ahmad Fadzil Mohd Hani, Aamir Saeed Malik, Raja Kamil, Ruslan Razak, and Azman Kiflie. Development of a non-invasive diagnostic tool for early detection of knee osteoarhritis. In *2011 National Postgraduate Conference*, pages 1–6. IEEE, 2011.

[20] CYJ Wenham, AJ Grainger, and PG Conaghan. The role of imaging modalities in the diagnosis, differential diagnosis and clinical assessment of peripheral joint osteoarthritis. *Osteoarthritis and cartilage*, 22(10):1692–1702, 2014.

[21] D. Hayashi, F.W. Roemer, and A. Guermazi. Imaging for osteoarthritis. *Annals of Physical and Rehabilitation Medicine*, 59(3):161 – 169, 2016.

[22] Nima Hafezi-Nejad, Ali Guermazi, Shadpour Demehri, and Frank W Roemer. New imaging modalities to predict and evaluate osteoarthritis progression. *Best Practice & Research Clinical Rheumatology*, 2018.

[23] JH Kellgren and JS Lawrence. Radiological assessment of osteoarthritis. *Annals of the rheumatic diseases*, 16(4):494, 1957.

[24] Mark D Kohn, Adam A Sassoon, and Navin D Fernando. Classifications in brief: Kellgren-lawrence classification of osteoarthritis, 2016.

[25] Mohit Kapoor and Nizar N Mahomed. *Osteoarthritis: Pathogenesis, diagnosis, available treatments, drug safety, regenerative and precision medicine*. Springer, 2015.

[26] Roy D Altman, James F Fries, Daniel A Bloch, John Carstens, T Cooke Derek Mb, Harry Genant, Philip Gofton, Harry Groth, Dennis J Mcshane, William A Murphy, et al. Radiographic assessment of progression in osteoarthritis. *Arthritis & Rheumatism*, 30(11):1214–1225, 1987.

[27] Lisa Sheehy and T Derek V Cooke. Radiographic assessment of leg alignment and grading of knee osteoarthritis: A critical review. *World Journal of Rheumatology*, 5(2):69–81, 2015.

[28] Joseph Antony, Kevin McGuinness, Kieran Moran, and 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*, pages 376–390. Springer, 2017.

[29] Tae Keun Yoo, Sung Kean Kim, Soo Beom Choi, Deog Young Kim, and Deok Won Kim. Interpretation of movement during stair ascent for predicting severity and prognosis of knee osteoarthritis in elderly women using support vector machine. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 192–196. IEEE, 2013.

[30] Amin Khatami, Abbas Khosravi, Thanh Nguyen, Chee Peng Lim, and Saeid Nahavandi. Medical image analysis using wavelet transform and deep belief networks. *Expert Systems with Applications*, 86:190–198, 2017.

[31] L Anifah, MH Purnomo, TLR Mengko, and IKE Purnama. Osteoarthritis severity determination using self organizing map based gabor kernel. In *IOP Conference Series: Materials Science and Engineering*, volume 306, page 012071. IOP Publishing, 2018.

[32] Rabia Riad, Rachid Jennane, Abdelbasset Brahim, Thomas Janvier, Hechmi Toumi, and Eric Lespessailles. Texture analysis using complex wavelet decomposition for knee osteoarthritis detection: Data from the osteoarthritis initiative. *Computers & Electrical Engineering*, 68:181–191, 2018.

[33] Luca Minciullo and Tim Cootes. Fully automated shape analysis for detection of osteoarthritis from lateral knee radiographs. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3787–3791. IEEE, 2016.

[34] Joseph Antony, Kevin McGuinness, Noel E O'Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1195–1200. IEEE, 2016.

[35] Berk Norman, Valentina Pedoia, Adam Noworolski, Thomas M Link, and Sharmila Majumdar. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *Journal of digital imaging*, pages 1–7, 2018.

[36] Aleksei Tiulpin, Stefan Klein, Sita Bierma-Zeinstra, Jérôme Thevenot, Esa Rahtu, Joyce van Meurs, Edwin HG Oei, and Simo Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *arXiv preprint arXiv:1904.06236*, 2019.

[37] 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.

[38] Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, and Khalid Ashraf. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*, 2017.

[39] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017.

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

[41] S Suresha, L Kidziński, E Halilaj, GE Gold, and SL Delp. Automated staging of knee osteoarthritis severity using deep neural networks. *Osteoarthritis and Cartilage*, 26:S441, 2018.

[42] Daichi Hayashi, Frank W Roemer, Mohamed Jarraya, and Ali Guermazi. Imaging in osteoarthritis. *Radiologic Clinics of North America*, 55(5):1085–1102, 2017.

[43] Ali Guermazi, Souhil Zaim, Bachir Taouli, Yves Miaux, Charles G Peterfy, and Harry K Genant. Mr findings in knee osteoarthritis. *European radiology*, 13(6):1370–1386, 2003.

[44] David W Stoller and Harry K Genant. Magnetic resonance imaging of the knee and hip. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 33(3):441–449, 1990.

[45] CG Peterfy, G Gold, F Eckstein, Flavia Cicuttini, B Dardzinski, and R Stevens. Mri protocols for whole-organ assessment of the knee in osteoarthritis. *Osteoarthritis and Cartilage*, 14:95–111, 2006.

[46] Gabrielle Blumenkrantz and Sharmila Majumdar. Quantitative magnetic resonance imaging of articular cartilage in osteoarthritis. *Eur Cell Mater*, 13(7), 2007.

[47] N Hafezi-Nejad, Shadpour Demehri, A Guermazi, and JA Carrino. Osteoarthritis year in review 2017: updates on imaging advancements. *Osteoarthritis and cartilage*, 26(3):341–349, 2018.

[48] John A Lynch, Frank W Roemer, Michael C Nevitt, David T Felson, Jingbo Niu, Charles B Eaton, and Ali Guermazi. Comparison of bloks and worms scoring systems part i. cross sectional comparison of methods to assess cartilage morphology, meniscal damage and bone marrow lesions on knee mri: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 18(11):1393–1401, 2010.

[49] Akash Gandhamal, Sanjay Talbar, Suhas Gajre, Ahmad Fadzil Mohd Hani, and Dileep Kumar. A generalized contrast enhancement approach for knee mr images. In *Signal and Information Processing (IConSIP), International Conference on*, pages 1–6. IEEE, 2016.

[50] Anthony Paproki, Craig Engstrom, Shekhar S Chandra, Ales Neubert, Jurgen Fripp, and Stuart Crozier. Automated segmentation and analysis of normal and osteoarthritic knee menisci from magnetic resonance images–data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 22(9):1259–1270, 2014.

[51] Brent Foster, Anand A Joshi, Marissa Borgese, Yasser Abdelhafez, Robert D Boutin, and Abhijit J Chaudhari. Wrist: A wrist image segmentation toolkit for carpal bone delineation from mri. *Computerized Medical Imaging and Graphics*, 63:31–40, 2018.

[52] Sanjeevakumar Kubakaddi, KM Ravikumar, and DG Harini. Measurement of cartilage thickness for early detection of knee osteoarthritis (koa). In *Point-of-Care Healthcare Technologies (PHT), 2013 IEEE*, pages 208–211. IEEE, 2013.

[53] Anthony Paproki, Katharine J Wilson, Rachel K Surowiec, Charles P Ho, Abinash Pant, Pierrick Bourgeat, Craig Engstrom, Stuart Crozier, and Jurgen Fripp. Automated segmentation and t2-mapping of the posterior cruciate ligament from mri of the knee: Data from the osteoarthritis initiative. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 424–427. IEEE, 2016.

[54] Gulshan B Sharma, Douglas D Robertson, Dawn A Laney, Michael J Gambello, and Michael Terk. Machine learning based analytics of micro-mri trabecular bone microarchitecture and texture in type 1 gaucher disease. *Journal of biomechanics*, 49(9):1961–1968, 2016.

[55] RP Dudhmande, AM Rajurkar, and VG Kottawar. Extraction of whole and torn meniscus in mri images and detection of meniscal tears. In *Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on*, pages 11–17. IEEE, 2017.

[56] Xu Dai and Janet Ronsky. The study of knee tibiofemoral condyle cartilage relaxation characters based on quantitative mr t2 imaging. In *Medical Imaging Physics and Engineering (ICMIPE), 2013 IEEE International Conference on*, pages 150–153. IEEE, 2013.

[57] C Burnett, P Wright, A-M Keenan, A Redmond, and J Ridgway. Magnetic resonance imaging of synovitis in knees of patients with osteoarthritis without injected contrast agents using t 1 quantification. *Radiography*, 2018.

[58] Chao Huang, Liang Shan, H Cecil Charles, Wolfgang Wirth, Marc Niethammer, and Hongtu Zhu. Diseased region detection of longitudinal knee magnetic resonance imaging data. *IEEE transactions on medical imaging*, 34(9):1914–1927, 2015.

[59] Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li, and Dimitris Metaxas. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 17–21. IEEE, 2015.

[60] Yukiko Yamamoto, Setsuo Tsuruta, Syoji Kobashi, Yoshitaka Sakurai, and Rainer Knauf. An efficient classification method for knee mr image segmentation. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2016 12th International Conference on*, pages 36–41. IEEE, 2016.

[61] AA Gatti. Neuralseg: state-of-the-art cartilage segmentation using deep learning–analyses of data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 26:S47–S48, 2018.

[62] Alexander Tack, Anirban Mukhopadhyay, and Stefan Zachow. Knee menisci segmentation using convolutional neural networks: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 26(5):680–688, 2018.

[63] Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.

[64] Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, and Christoph Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *International MICCAI Brainlesion Workshop*, pages 241–252. Springer, 2017.

[65] ST Soellner, A Goldmann, D Muelheims, GH Welsch, and ML Pachowsky. Intraoperative validation of quantitative t2 mapping in patients with articular cartilage lesions of the knee. *Osteoarthritis and cartilage*, 25(11):1841–1849, 2017.

[66] Kathy Zheng, Scott Martin, Christopher Rashidifard, Bin Liu, and Mark E Brezinski. In vivo micron scale arthroscopic imaging of human knee osteoarthritis with oct: Comparison with mri and arthroscopy. In *Conference on Lasers and Electro-Optics*, page CMCC3. Optical Society of America, 2009.

[67] Mahesh B Nagarajan, Paola Coan, Markus B Huber, Paul C Diemoz, Christian Glaser, and Axel Wismüller. Computer-aided diagnosis in phase contrast imaging x-ray computed tomography for quantitative characterization of ex vivo human patellar cartilage. *IEEE Transactions on Biomedical Engineering*, 60(10):2896–2903, 2013.

[68] Anas Z Abidin, Botao Deng, Adora M DSouza, Mahesh B Nagarajan, Paola Coan, and Axel Wismüller. Deep transfer learning for characterizing chondrocyte patterns in phase contrast x-ray computed tomography images of the human patellar cartilage. *Computers in biology and medicine*, 95:24–33, 2018.

[69] Priscille de Dumast, Clément Mirabel, Lucia Cevidanes, Antonio Ruellas, Marilia Yatabe, Marcos Ioshida, Nina Tubau Ribera, Loic Michoud, Liliane Gomes, Chao Huang, et al. A web-based system for neural network based classification in temporomandibular joint osteoarthritis. *Computerized Medical Imaging and Graphics*, 67:45–54, 2018.

[70] Prajna Ramesh Desai and Ilker Hacihaliloglu. Enhancement and automated segmentation of ultrasound knee cartilage for early diagnosis of knee osteoarthritis. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1471–1474. IEEE, 2018.

[71] Md Belayet Hossain, Khin Wee Lai, Belinda Pingguan-Murphy, Yan Chai Hum, Maheza Irna Mohd Salim, and Yih Miin Liew. Contrast enhancement of ultrasound imaging of the knee joint cartilage for early detection of knee osteoarthritis. *Biomedical Signal Processing and Control*, 13:157–167, 2014.

[72] Jana Podlipská, Ali Guermazi, Petri Lehenkari, Jaakko Niinimäki, Frank W Roemer, Jari P Arokoski, Päivi Kaukinen, Esa Liukkonen, Eveliina Lammentausta, Miika T Nieminen, et al. Comparison of diagnostic performance of semi-quantitative knee ultrasound and knee radiography with mri: Oulu knee osteoarthritis study. *Scientific reports*, 6:22365, 2016.

[73] Krzysztof Kręcisz and Dawid Bączkowicz. Analysis and multiclass classification of pathological knee joints using vibroarthrographic signals. *Computer methods and programs in biomedicine*, 154:37–44, 2018.

[74] Xiao Wang, Ming-Yang Zhai, Zhi-Hua Mao, Yan-Fei Lu, and Jian-Hua Yin. Fourier transform infrared spectroscopic imaging application for multi-stage discrimination in cartilage degeneration. *Infrared Physics & Technology*, 2018.

[75] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[76] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[77] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 3(3):173, 2019.

[78] Kambam Bijen Singh, Telajala Venkata Mahendra, Ravi Singh Kurmvanshi, and Ch V Rama Rao. Image enhancement with the application of local and global enhancement methods for dark images. In *2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*, pages 199–202. IEEE, 2017.

[79] Sampada S Pathak, Prashant Dahiwale, and Ganesh Padole. A combined effect of local and global method for contrast image enhancement. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 1–5. IEEE, 2015.

[80] Elena S Yelmanova and Yuriy M Romanyshyn. Automatic histogram-based contrast enhancement for low-contrast images with small-sized objects. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 965–970. IEEE, 2017.

[81] Sergei Yelmanov and Yuriy Romanyshyn. Automatic enhancement of low-contrast monochrome images. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, pages 587–593. IEEE, 2018.

[82] Tarun Kumar Agarwal, Mayank Tiwari, and Subir Singh Lamba. Modified histogram based contrast enhancement using homomorphic filtering for medical images. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 964–968. IEEE, 2014.

[83] Randeep Kaur and Sandeep Kaur. Comparison of contrast enhancement techniques for medical image. In *2016 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pages 155–159. IEEE, 2016.

[84] Soong-Der Chen and Abd Rahman Ramli. Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation. *IEEE Transactions on consumer Electronics*, 49(4):1301–1309, 2003.

[85] SS Bedi and Rati Khandelwal. Various image enhancement techniques-a critical review. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(3), 2013.

[86] Andrea Polesel, Giovanni Ramponi, and V John Mathews. Image enhancement via adaptive unsharp masking. *IEEE transactions on image processing*, 9(3):505–510, 2000.

[87] Abdalla Mohamed Hambal, Zhijun Pei, and Faustini Libent Ishabailu. Image noise reduction and filtering techniques. *International Journal of Science and Research (IJSR)*, 6(3):2033–2038, 2017.

[88] V Kamalaveni, R Anitha Rajalakshmi, and KA Narayanankutty. Image denoising using variations of perona-malik model with different edge stopping functions. *Procedia Computer Science*, 58:673–682, 2015.

[89] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7):629–639, 1990.

[90] Michael J Black, Guillermo Sapiro, David H Marimont, and David Heeger. Robust anisotropic diffusion. *IEEE Transactions on image processing*, 7(3):421–432, 1998.

[91] Hassan Masoumi, Alireza Behrad, Mohammad Ali Pourmina, and Alireza Roosta. Automatic liver segmentation in mri images using an iterative watershed algorithm and artificial neural network. *Biomedical signal processing and control*, 7(5):429–437, 2012.

[92] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[93] Phi Vu Tran. A fully convolutional neural network for cardiac segmentation in short-axis mri. *arXiv preprint arXiv:1604.00494*, 2016.

[94] 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.

[95] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[96] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. In *Handbook of Deep Learning Applications*, pages 161–200. Springer, 2019.

[97] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[98] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

[99] 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.

[100] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[101] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[102] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[103] Shaohua Wan, Zhijun Chen, Tao Zhang, Bo Zhang, and Kong-kat Wong. Bootstrapping face detection with hard negative examples. *arXiv preprint arXiv:1608.02236*, 2016.

[104] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[105] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[106] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[107] Neil A Segal, Michael C Nevitt, K Douglas Gross, Jean Hietpas, Natalie A Glass, Cora E Lewis, and James C Torner. The multicenter osteoarthritis study (most): Opportunities for rehabilitation research. *PM & R: the journal of injury, function, and rehabilitation*, 5(8), 2013.

[108] Alireza Baratloo, Mostafa Hosseini, Ahmed Negida, and Gehad El Ashal. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. 2015.

[109] Jaime S Cardoso and Ricardo Sousa. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08):1173–1195, 2011.

[110] Lisa Gaudette and Nathalie Japkowicz. Evaluation methods for ordinal classification. In *Canadian Conference on Artificial Intelligence*, pages 207–210. Springer, 2009.

Typeset September 26, 2022