Deep learning in the detection of early inflammatory signs in rheumatoid arthritis

Yanli Li

Colophon

About the cover:
The cover will be created by Yanli Li.

Deep learning in the detection of early inflammatory signs in rheumatoid arthritis
Yanli Li

ISBN: undetermined
Thesis layout & cover designed by Yanli Li
Printed by Ridderprint, the Netherlands

Deep learning in the detection of early inflammatory signs in rheumatoid arthritis

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus undetermined,
volgens besluit van het college voor promoties
te verdedigen op 2025
klokke undetermineduur

door

Yanli Li
geboren te Chengdu, Sichuan, China
in 1995

Promotor:	Prof. dr. ir. M. Staring
Co-promotor:	Dr. B. C. Stoel
Leden promotiecommissie:	Prof. dr. A.H.M. van der Helm-van Mil Prof. dr. H. A. Marquering *Amsterdam University Medical Center, Amsterdam* Prof. dr. Katherine Breininger *Universitat Wurzburg, Würzburg* dr. Jifke Veenland *Erasmus Medical Center, Erasmus*

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands.

Financial support for the publication of this thesis was kindly provided by:
LKEB,
Library of Leiden University

Contents

List of abbreviations **v**

1 Introduction **1**
1.1 Rheumatoid arthritis 1
1.2 Wrist, MCP and MTP joints 3
1.2.1 Anatomy of the Wrist, MCP and MTP joints 3
1.2.2 MRI of the wrist, MCP and MTP joints 5
1.3 Inflammation assessment for RA 7
1.4 Deep learning in inflammation assessment and RA 8
1.5 Explainable deep learning methods 9
1.6 Thesis overview 11

2 Automatic joint inflammation estimation based on regression neural networks **15**
2.1 Introduction 17
2.2 Methods 21
2.2.1 Data 21
2.2.2 Preprocessing 23
2.2.3 Model architecture and training 25
2.2.4 Testing, validation and statistical analysis 25
2.2.5 Reliability and explainability 27
2.3 Results 27
2.3.1 Performance of Route 1 and comparison with human experts . . 28
2.3.2 Performance of Route 2 and comparison with human experts . . 29
2.3.3 Output distribution of ADMIRA inflammation assessment 30
2.3.4 Comparison with existing methods 30
2.3.5 Explainability through CAM 31
2.4 Discussion 32
2.4.1 Data distribution 32

2.4.2 Limitations . . . 34
2.4.3 The two routes for inflammation estimation . . . 37
2.4.4 Summary . . . 37
2.5 Conclusion . . . 37

3 Rheumatoid arthritis classification and prediction by consistency-based deep learning using extremity MRI scans 39
3.1 Introduction . . . 41
3.2 Materials . . . 42
3.2.1 Structure of materials . . . 42
3.2.2 Task definition . . . 44
3.3 Methods . . . 44
3.3.1 Overall workflow . . . 44
3.3.2 Preprocessing . . . 45
3.3.3 Backbone model . . . 46
3.3.4 Anatomical consistency and self-supervised pretraining . . . 47
3.3.5 Label-consistency loss function . . . 49
3.3.6 Class activation mapping . . . 50
3.4 Results . . . 51
3.4.1 Evaluation principles . . . 51
3.4.2 Reconstruction examples from self-supervised pretraining . . . 51
3.4.3 Overall performance on the four tasks . . . 51
3.4.4 General improvements compared to baseline models . . . 53
3.4.5 Ablation study of each proposed component . . . 56
3.4.6 Saliency maps generated by CAM . . . 56
3.5 Discussion . . . 56
3.6 Conclusions . . . 60

4 Feature analysis for proper intensity scaling and feature distinction in class activation maps 61
4.1 Introduction . . . 63
4.2 Method . . . 68
4.2.1 Global intensity scale . . . 68
4.2.2 Feature distinction . . . 71
4.2.3 Evaluation: global intensity scale . . . 72
4.2.4 Evaluation: feature distinction . . . 73
4.3 Materials . . . 75
4.3.1 Datasets and models . . . 75
4.3.2 CAM algorithms . . . 77

4.4 Experiments and results . . . 77
4.4.1 Global intensity scaling . . . 78
4.4.2 Feature distinction using importance matrix . . . 79
4.4.3 Visual examples . . . 83
4.5 Discussion . . . 86
4.5.1 Uses of the feature distinction . . . 87
4.5.2 Definition of the importance . . . 89
4.5.3 Why are some CAM algorithms not included in the experiments? 89
4.5.4 Metrics like Dice or IoU to evaluate CAM algorithms? . . . 89
4.5.5 Generalizability of global intensity scaling . . . 90
4.5.6 Generalizability of the feature analysis . . . 90
4.5.7 Hyper-parameter issues . . . 90
4.5.8 Other topics . . . 91
4.6 Conclusion . . . 91

5 Aggregation of Class Activation Maps for Explaining Deep Learning at a Population Level 93
5.1 Introduction . . . 95
5.2 Method . . . 97
5.2.1 Class activation mapping with a global intensity scale . . . 97
5.2.2 Region definition through segmentation . . . 98
5.2.3 Aggregation process . . . 99
5.2.4 The conceptual method for validating the aggregation . . . 99
5.2.5 The simulation for validating the aggregation . . . 101
5.3 Datasets . . . 103
5.3.1 Simulation 1: qualitative evaluation . . . 103
5.3.2 Simulation 2: quantitative evaluation . . . 105
5.3.3 Rheumatoid arthritis prediction . . . 105
5.3.4 Stenosis score prediction task . . . 106
5.4 Experiments and results . . . 107
5.4.1 Qualitative validation on the Simulation 1 . . . 107
5.4.2 Quantitative validation on the Simulation 2 . . . 108
5.4.3 Application to rheumatoid arthritis prediction task . . . 109
5.4.4 Application to predicting stenosis score . . . 110
5.5 Discussion . . . 111
5.5.1 Bias of models and datasets . . . 111
5.5.2 Accuracy of segmentation . . . 111
5.5.3 Definition of importance . . . 111
5.5.4 CAM aggregation: a qualitative evaluation . . . 112

5.6 Conclusion . . . 112

6 Auxiliary-branch CAM in deep learning models serves as a tool for discovering undefined image patterns of rheumatoid arthritis 113
6.1 Introduction . . . 115
6.2 Material and method . . . 117
6.3 Material . . . 117
6.3.1 General workflow . . . 118
6.3.2 Model architecture . . . 119
6.3.3 Class activation mapping and aggregation . . . 120
6.3.4 Validation method . . . 120
6.4 Experiments . . . 121
6.4.1 Backbone model performance . . . 121
6.4.2 Validation of the framework . . . 121
6.5 Discussion . . . 123
6.6 Conclusion . . . 124
6.7 Acknowledgments . . . 125

7 Summary, discussion and future work 127
7.1 Summary . . . 127
7.2 Discussion on limitations and future work . . . 129
7.2.1 Discussion per topic from each chapter . . . 129
7.2.2 General discussion . . . 131
7.3 General conclusions . . . 134

8 Samenvatting, discussie en toekomstig werk 135
8.1 Samenvatting . . . 135
8.2 Discussie over beperkingen en toekomstig werk . . . 137
8.2.1 Discussie binnen elk hoofdstuk . . . 137
8.2.2 Algemene discussie . . . 139
8.3 Algemene conclusies . . . 142

References 143

List of publications 159

Acknowledgements 161

Curriculum Vitae 163

List of abbreviations

DL	deep learning	1
AI	artificial intelligence	8
CNNs	convolution neural networks	44
MRI	magnetic resonance imaging	1
CAM	class activation mapping	10
CAMs	class activation maps	11
RA	rheumatoid arthritis	1
EAC	recent-onset arthritis	43
CSA	clinically suspect arthralgia	3
BME	bone marrow edema	3
TSY	tenosynovitis	3
SYN	synovitis	3
TRA	transversal	6
COR	coronal	5
RAMRIS	Rheumatoid Arthritis Magnetic Resonance Imaging Scoring	8
MCP	metacarpophalangeal	3
MTP	metatarsophalangeal	3
R	Pearson's correlation coefficient	13
MSE	mean squared error	48
SD	standard deviation	51
ROI	regions of interest	8
ADMIRA	automatic DL-based MRI analysis of inflammatory signs in RA	12

1

Introduction

This thesis focuses on detecting inflammatory signs in Rheumatoid arthritis (rheumatoid arthritis (RA)) and predicting RA development, based on magnetic resonance imaging (MRI) using explainable deep learning (DL) models. This introduction successively describes the following topics:

- The medical background of RA, the value of detecting early signs of RA, the advantages of image-based detection and the necessity of automatic detection.
- The anatomical structures typically affected by RA, the reasons of using MRI as the image modality for detection, and manually-defined image biomarkers in these MRI images.
- The prevailing methods for detecting early signs and predicting RA development based on MRI scans, the disadvantages and current solutions.
- New possibilities after the development of DL and the challenges of applying DL in the field.
- DL's explainability and related problems.
- Overview of the different chapters in the thesis.

1.1 Rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disorder that especially affects joints in the wrists, hands and feet [3, 4]. It can affect the synovial lining of joints, causing a painful swelling that can eventually result in bone erosions and joint deformities. Furthermore, it may not only affect joints, but in some cases also a wide variety of body systems, including the skin, eyes, lungs, heart and blood vessels. Based on the Global Burden of Disease (GBD) study, RA affected 17.6 million people worldwide in 2020, thereby remaining one of the major causes of disability and labor loss [5].

At present, RA cannot be cured, but it can frequently be controlled through treatment. When RA has already developed, treatment typically aims at reducing

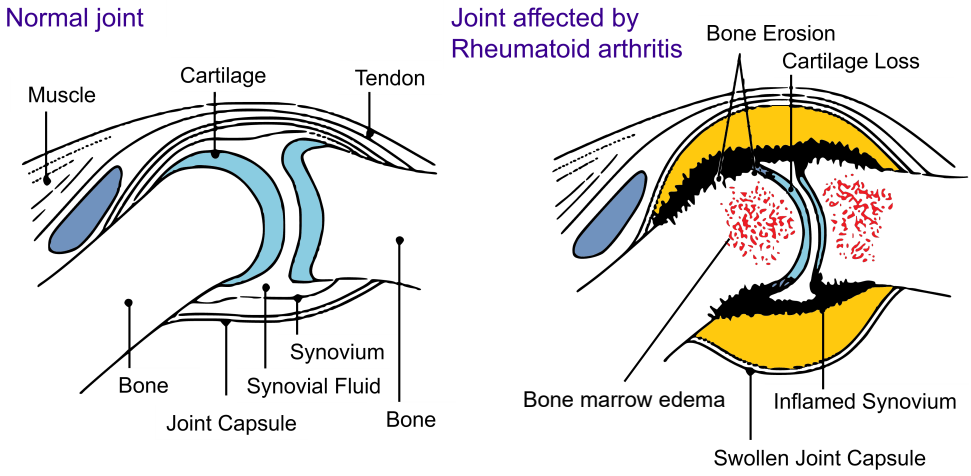


Figure 1.1: Depiction of the anatomy of a normal synovial joint and pathology of a joint affected by rheumatoid arthritis. The synovial joint is the most common type of joint in the human body, which consists of two bones covered with articular cartilage, synovial fluid that connects bones and synovial membrane (synovium) that encapsulates bone surfaces and the fluid. RA primarily manifests itself as inflammation of the joints [1]. (Adapted from Wikimedia [1, 2])

inflammation, relieving symptoms, preventing joint or organ damage, and slowing RA progression. Fortunately, the clinical outcomes have been improved, mainly because the strategy shifts from slowing joint damage to preventing damage before it occurs [6, 7, 8]. This shift relies heavily on early initiation of disease-modifying antirheumatic drugs and early diagnosis, underscoring the importance of early identification of inflammatory signs and even prediction of RA development.

The method for detecting inflammatory signs at an early stage is one of the most important research topics in RA. Many approaches for diagnosis and prediction are based on serum biomarkers and have been widely-recognized, such as Rheumatoid Factor (RF), Auto-antibodies Against Citrullinated Proteins (ACPAs) and Erythrocyte Sedimentation Rate (ESR) and C-reactive Protein (CRP). While these serum biomarkers are successful, alternative methods based on imaging modalities have their advantages.

Compared to serum biomarkers, images can provide more specific anatomical locations of the inflammatory signs, contribute to detecting progress and treatment, and serve as diagnostic basis when serological tests are atypical (20%-30% atypical when diagnosing RA [9]).

Based on this motivation, researchers started their investigations in detecting early RA signs and predicting future RA development by defining and assessing image biomarkers. Fig. 1.1 presents a depiction of a normal synovial joint and pathology of

a joint affected by rheumatoid arthritis. Researchers assess the inflammation within this kind of joints by visually inspect images, define specific inflammation as image biomarkers, and build criteria-based systems to quantify the severity of these image biomarkers. Similarly, trained readers could evaluate tendon sheaths, synovia, and bones, quantifying the thickness of peritendinous effusion, synovial effusion and synovial proliferation, and the volume of bone marrow edema respectively. By this kind of quantification of inflammation severity in specific regions, researchers can detect and predict future RA development to some extent [10].

However, this is a laborious and difficult task that requires significant time investment, is prone to inter- and intra- reader variability and demands rigorous training to ensure accuracy. The automation of assessing such inflammation in images is therefore essential to facilitate the detection and prediction of RA at an early stage [11].

In this study, we adopted medical images of patients from clinically suspect arthralgia (CSA) group [12, 13] that are considered to be an at-risk stage of RA, and images of other groups related to undifferentiated or rheumatoid arthritis as control group. While some CSA-patients may not develop clinical arthritis, a certain proportion of CSA-patients without preventive treatment could develop RA and the inflammatory signs are expected to be detectable during this period. Taking advantages of manually-defined inflammatory signs and clinical records of RA development of these CSA patients, this study aims at determining whether the detection of early RA signs and RA prediction based on images can be automated through a fully data-driven method.

1.2 Wrist, MCP and MTP joints

1.2.1 Anatomy of the Wrist, MCP and MTP joints

As introduced, RA especially affects joints in the wrists, hands and feet [3, 4]. Specifically, the synovitis (SYN), tenosynovitis (TSY) and bone marrow edema (BME) in wrist, metacarpophalangeal (MCP) and metatarsophalangeal (MTP) joints are typically considered to be potentially predictive to RA development. These anatomical structures have however complex shapes with complicated three-dimensional interconnections, as can be seen in Fig. 1.2.

The wrist is composed of bones, joints, tendons and blood vessels, which together provide a wide range of motion and mechanical support for hand function. The wrist contains eight carpal bones (arranged in a proximal and distal row) and three primary joints (the radiocarpal joint, midcarpal joint and the distal radioulnar joint) [14]. The wrist is traversed by several tendons that control hand and finger movement, six main tendons on the palmar side and six compartments beneath the extensor retinaculum on the dorsal side [15]. These tendons are surrounded by synovial sheaths, which reduce

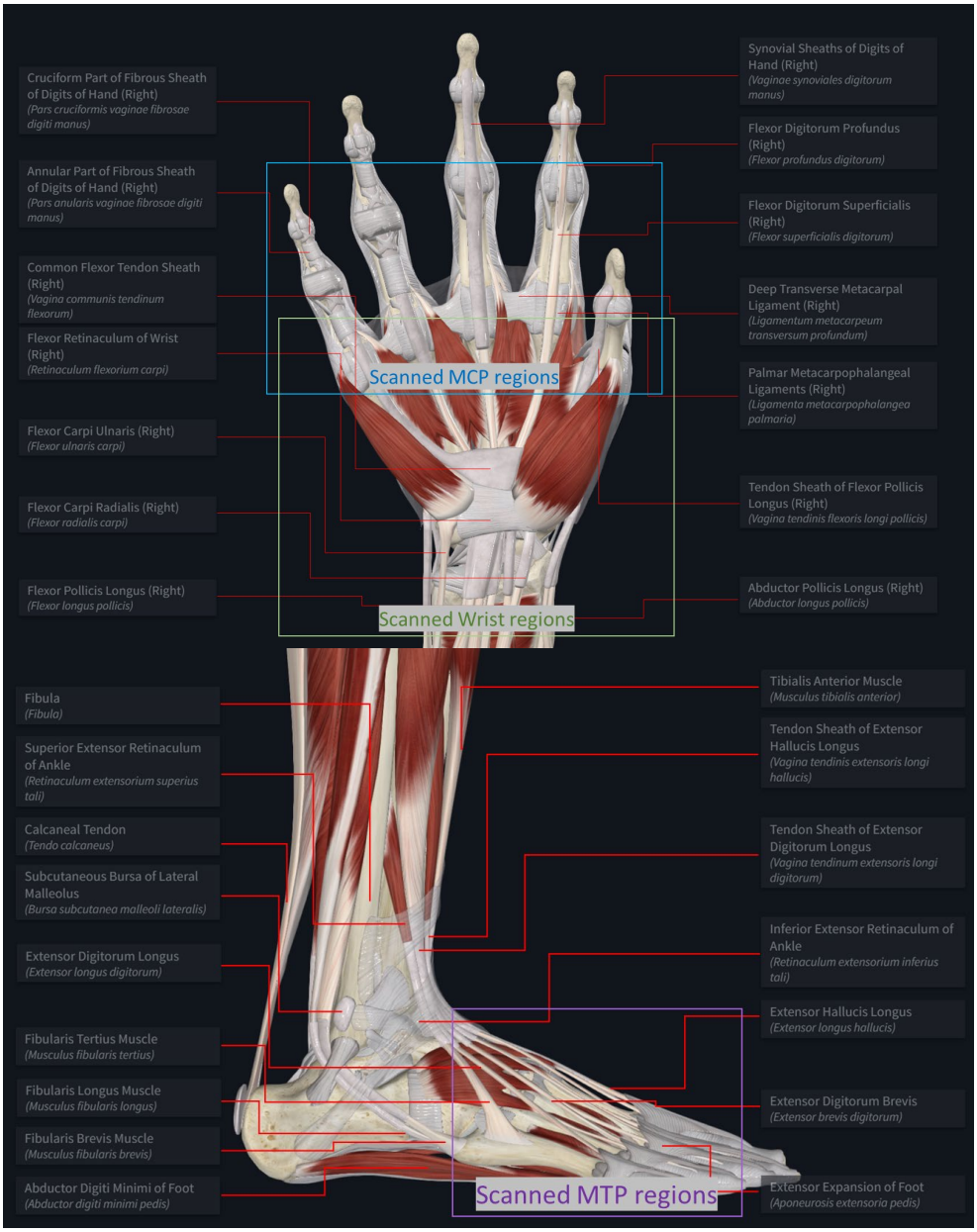


Figure 1.2: Anatomy for the tendon (sheaths) of palmar-viewed hand (including the wrist and MCP joints) and lateral-viewed feet (including MTP joints). Green, blue and purple boxes highlight the anatomical regions scanned by MRI that were used in this study. (Images adapted from `https://3d4medical.com/`)

friction and facilitate smooth motion. The carpal tunnel and extensor compartments contain synovial sheaths that enclose the tendons. Additionally, synovial membranes line the internal surfaces of the wrist joints, secreting synovial fluid for lubrication. Bursae, such as the ulnar and radial bursae, serve to cushion tendons during movement [16].

The MCP joint is a synovial, condyloid joint formed between the heads of the metacarpal bones and the bases of the proximal phalanges [14]. Each MCP joint is formed by the articulation between the convex head of a metacarpal bone and the concave base of a proximal phalanx 1.1. The tendons of MCP joints contain the flexor and extensor tendons for finger flexion and extension, respectively. As for the synovial structures, synovial membranes in the joint capsules and synovial sheaths around the flexor tendons are the main synovial structures in the MCP joint.

The MTP joint is located between the metatarsal bones and the proximal phalanges of the toes and play a central role in the biomechanics of walking and running. Each MTP joint is formed by the articulation of the rounded head of a metatarsal bone with the shallow base of a proximal phalanx. Similar to MCP joints, MTP joints include flexor and extensor tendons. The MTP joint is lined by a synovial membrane and surrounded by a fibrous capsule. The tendons of the flexor and extensor muscles are encased in synovial sheaths.

In this thesis, bones, synovial sheaths around the tendons, the synovial membrane (synovium) between bones are considered the major regions of interests, as these regions can contain inflammatory signs that are predictive of RA, according to clinical studies [17, 18].

1.2.2 MRI of the wrist, MCP and MTP joints

Among imaging modalities, MRI is the most sensitive to detect inflammation of tendon sheaths, synovium, and bones, and has therefore become a promising method to detect subtle inflammatory signs that are predictive of RA [24]. In this study, MRI is therefore the main image modality for most of the chapters.

Some examples of the MRI scans of the wrist, MCP and MTP joints used in this study are shown in Fig. 1.3, including vessels (indicated by blue lines), tendons (green lines), bones (orange lines) and synovia (red lines). As can be seen in this figure, the bones, tendons and synovia present medium signal intensities, and vessels present high signal intensities.

These MRI examples are acquired with a 1.5T extremity MRI scanner (GE Healthcare) using a 100-mm coil, with contrast enhancement (T1-Gd) and frequency-selective fat saturation. For the coronal (COR) scans (3D scans with the highest resolution in the coronal plane), the repetition time was 650 ms, echo time 17 ms, acquisition

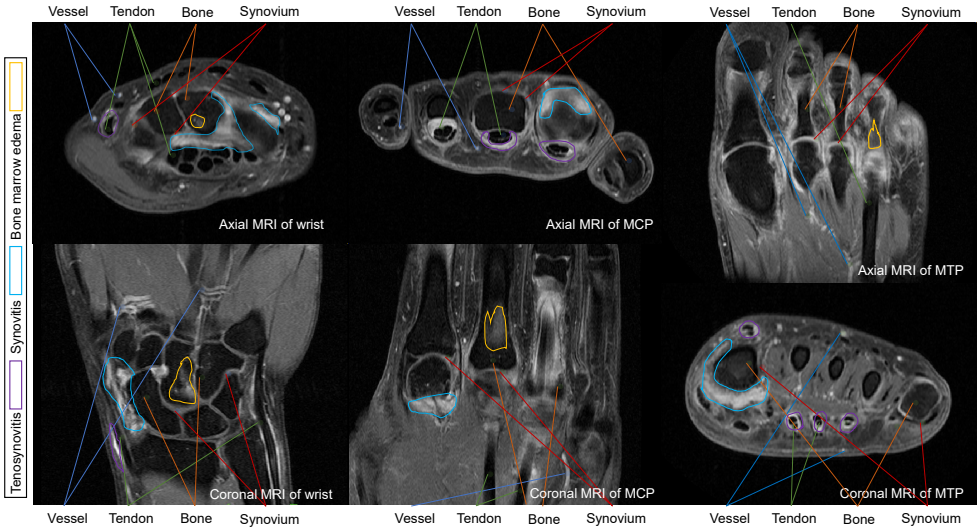


Figure 1.3: MRI examples of anatomical structures of interest and the inflammatory signs in the wrists (hand), MCPs (hand) and MTPs (feet) relevant to this study, partly adopted from previous clinical study on the same topic [19, 20, 21]. Inflammatory signs, SYN, TSY and BME, were defined under the guidance of [22, 23]. In these examples, regions around the tendon sheaths, bones and synovia are the most typical regions of inflammatory signs. Some of these anatomical structures are indicated by green, orange and red lines. Some vessels with high intensities are indicated by blue lines. Some of the regions with purple, light blue and light red contours represent the inflammatory signs of SYN, TSY and BME, respectively.

matrix 364x224, echo train length 2, slice thickness 2mm, and slice gap 0.2 mm. For transversal (TRA) scans, these parameters are: 570 ms, 7 ms, 320x192), 2, 3 mm, and 0.3 mm, respectively [25].

Fig. 1.3 also shows MRI examples of joints that are typically affected by RA, illustrating major anatomical structures and some regions of manually-defined inflammatory signs – TSY, SYN and BME [17, 18].

Tenosynovitis refers to the inflammation of the tendon sheath, particularly affecting the synovial lining that surrounds tendons. It is a hallmark of several inflammatory arthritides, including RA. In T1-Gd MRI scans, tenosynovitis presents as fluid or thickening within the tendon sheath, with high signal intensities after injection of a gadolinium contrast agent [26]. Tenosynovitis most commonly affects the extensor and flexor tendons in the wrist [27], flexor tendon sheath of the MCP joint and less frequently observed in the MTP joint [28].

Synovitis denotes the inflammation of the synovial membrane, leading to synovial thickening and joint effusion. Synovitis is an important inflammatory sign for both di-

agnostic and disease monitoring of autoimmune diseases such as RA [29]. Synovitis in the wrist involves the radiocarpal, intercarpal-carpometacarpal, and distal radioulnar joints. Together with synovitis of the MCP joints, these are classic inflammatory signs of RA [30]. MRI can reveal erosions associated with chronic synovial inflammation [28].

BME is characterized by increased water content within the trabecular bone. BME typically indicates active inflammation and correlates with poor outcomes in RA [28].

These inflammatory signs that are (potentially) predictive of RA can be effectively visualized through MRI, and are detectable after fat suppression and contrast enhancement. On T1-Gd MRI scans, tenosynovitis and synovitis appear as areas of high signal intensities, BME shows mildly to moderately increased signal intensities.

1.3 Inflammation assessment for RA

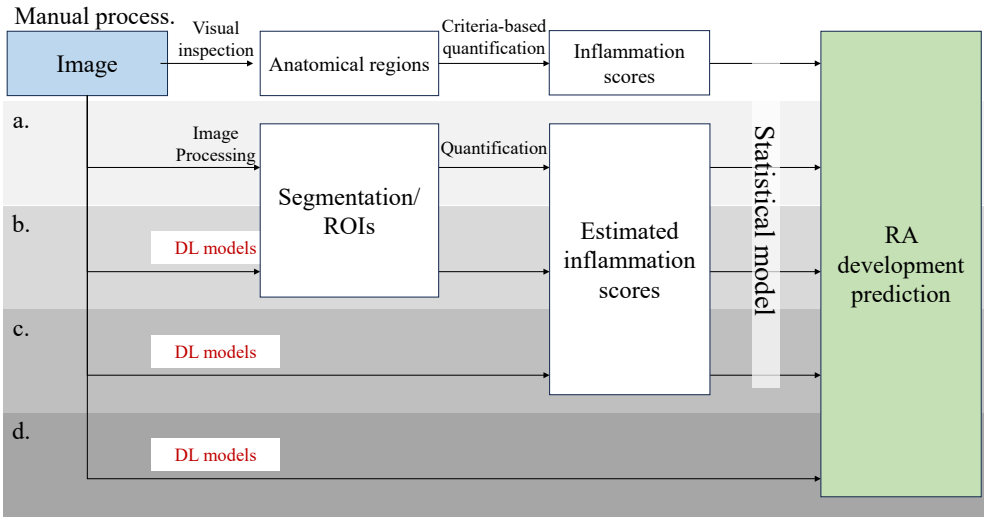


Figure 1.4: Different routes to predict future RA development or other outcomes. (a) A route of classical segmentation with quantification to obtain inflammation assessment, and apply statistical models developed by rheumatologists to obtain diagnosis and prediction; (b) DL-based segmentation with classical quantification, then apply statistical models; (c) DL-based inflammation assessment plus statistical models; (d) End-to-end DL models that directly output diagnosis or prediction. The grayscale of the rectangular box represents the degree of artificial intelligence of the route.

To detect early signs or predict RA based on MRI, the first step is to detect and quantify early inflammatory signs in these MRI scans, specifically the highlighted inflammatory signs in Fig. 1.3. The prevailing methods and golden standards of inflammation assessment based on MRI scans are visual scoring systems such as

the Rheumatoid Arthritis Magnetic Resonance Imaging Scoring (RAMRIS) system [31]. Using these scoring systems, trained readers evaluate multiple tendon sheaths, synovia and bones. In this way, these scoring systems provide (semi-)quantitative measurements of multiple image biomarkers, including those describing inflammation severity. Then rheumatologists can apply their statistical models based on specific variables (combinations of these inflammatory signs) for analyzing the development of certain disease with high accuracy, such as early RA in this specific study [10]. This process is illustrated in the top row of Fig. 1.4.

This visual inflammation assessment however involves tedious work, rigorous training for assessing joint inflammation and implicit inter-/intra-reader disagreement. This triggers studies in automatic quantification of joint inflammation. Many studies followed a segmentation-quantification approach [32, 33, 34] to quantify synovitis [35], tenosynovitis and bone marrow edema (BME) [36, 25] (see Fig. 1.4 a), achieving promising results, provided an available accurate segmentation or regions of interest (ROI). This segmentation-quantification flow is intuitive and explainable, and is therefore the previously prevailing method for automatically assessing joint inflammation in RA.

1.4 Deep learning in inflammation assessment and RA

With the development of artificial intelligence (AI) methods especially deep learning, more accurate segmentation methods [37] were developed for this segmentation-quantification framework (see Fig. 1.4 b). However, the ground truth of the segmentations requires tedious work of experts and massive time, sometimes the cost is even higher than having manual inflammation assessment. Meanwhile, automatic joint inflammation assessment without the need for segmentation or ROIs became a possible route as DL models are capable of assessing such image biomarkers in other medical tasks [38, 39, 40, 41, 42, 43, 44, 45, 46] (see Fig. 1.4 c). Furthermore, a new route has emerged – using deep learning to achieve fully data-driven direct diagnosis or prediction of RA based on the original images, namely an end-to-end method (see Fig. 1.4 d), which is capable of providing a new perspective independent of expert knowledge and potentially provides additional clues on predictive inflammatory signs.

This kind of end-to-end methods that take in original images and output corresponding labels, using a large amount of data to automatically find the disease-relevant image patterns, have been investigated in many other medical tasks [47, 48, 49]. However, these have not been explored systematically in diagnosing or predicting RA. The challenges to be overcome for training DL models of detecting and predicting (early) RA are as follows:

- Limited models or methods are built for the problem of detecting or predicting early RA.

- The time window, for collecting iamges with early RA signs is narrow. It requires collecting images from arthralgia (potential early RA symptom) patients with possible early signs of RA. This complicates data acquisition and results in a limited dataset size. Consequently, overfitting becomes a severe problem, and the small size of the dataset also restricts the performances of large models with massive trainable parameters.
- The variety and complexity of anatomical and pathological structures in the hands and feet, and variability in positioning the hands and feet, further amplifies the difficulty of the tasks, resulting in insufficient performances of standard models for medical images.
- Artifacts caused by fat suppression errors, movement or aliasing may significantly influence the automatic interpretation of MRI scans.
- Lack of publicly accessible datasets with similar subjects or tasks. DL models failed to benefit from transfer learning and well-developed preprocessing methods in pre-experiments.
- Lastly, DL models need to be explainable, so that rheumatologist can be convinced that the model is accurate and trustworthy for clinical use in the future, as elaborated in the next section.

1.5 Explainable deep learning methods

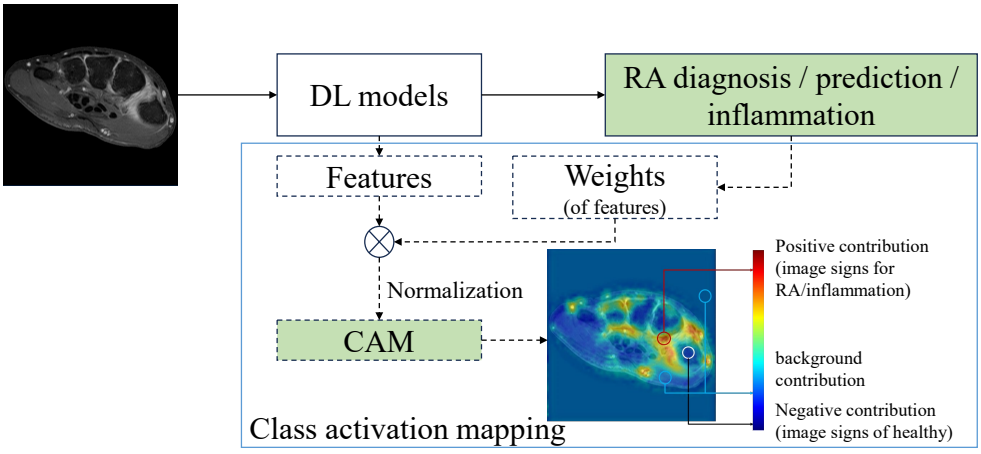


Figure 1.5: The calculation of a class activation map (CAM) that highlights the focus of a deep learning model.

The methods based on the route segmentation-quantification are intuitive and explainable because each step in these methods can be evaluated and corrected. But

this goes at a cost of implicit upper limits from the manually-defined biomarkers, as the manually-defined biomarkers may fail to include all relevant signals in MRI scans. On the contrary, the end-to-end methods based on DL models may be capable of discovering some new image biomarkers, yet they sacrifice the explainability and face overfitting on confounders. For example, when an end-to-end approach goes wrong during diagnosing specific disease based on symptoms at specific regions, it is challenging to figure out whether the error comes from the errors of the implicit segmentation or not, and it is even challenging to ensure the model has an implicit segmentation.

The great success of DL in some medical tasks [47, 48, 49] relies on a large amount of data to train and validate the DL models, which is not feasible for most datasets. The lack of enough data for training not only limits the accuracy of DL-based methods, but also places a higher risk of overfitting and learning confounding patterns.

A solution to this risk, in addition to collecting large-scale datasets for validation, is DL explainability. It needs to be confirmed that DL models are actually reasoning based on trustworthy evidence and convincing features [50, 51, 52, 53, 54, 55, 56]. In medical image analysis, this confirmation specifically refers to ensuring that the focus of a model on the input images should be spatially and semantically consistent with the focus of clinical experts [57].

For DL, the most popular method is the family of class activation mapping (CAM) algorithms [55, 58, 59], which reveals the focus of DL models by generating attention maps (also known as heatmaps or saliency maps), where the signal intensities represent the importance (class specificity or contribution) of a region to the output of the DL model (see Fig. 1.5). By comparing these focuses and expert knowledge for all inputs in the dataset one by one, the alignment between DL models and expert knowledge can be verified and possible disagreements can be located. This method provides an approach to reveal DL models' reasoning when diagnosing or predicting RA, and the inflammatory signs that indicate early RA could be located by checking the highlighted regions in the attention maps.

While explainable DL using CAMs has been investigated in many natural imaging fields and only a few medical imaging fields, two important aspects of CAMs remain unexplored and are topics in this thesis:

- Intensity scaling for normalizing the CAMs to correctly highlight ROIs and avoid misinterpretation. (Chapter 4)
- Population-level analysis after obtaining individual CAMs to achieve conclusions on the relationship between specific regions and outputs. (Chapter 5)

The improper intensity scaling could very likely lead to misinterpretation of CAMs and consequently affect the correctness of the conclusions. The lack of population-level

analysis limits the use of CAMs and DL methods to merely an individual visual check - any population-level conclusions would require reader studies and again become tedious, subjective and time-consuming.

1.6 Thesis overview

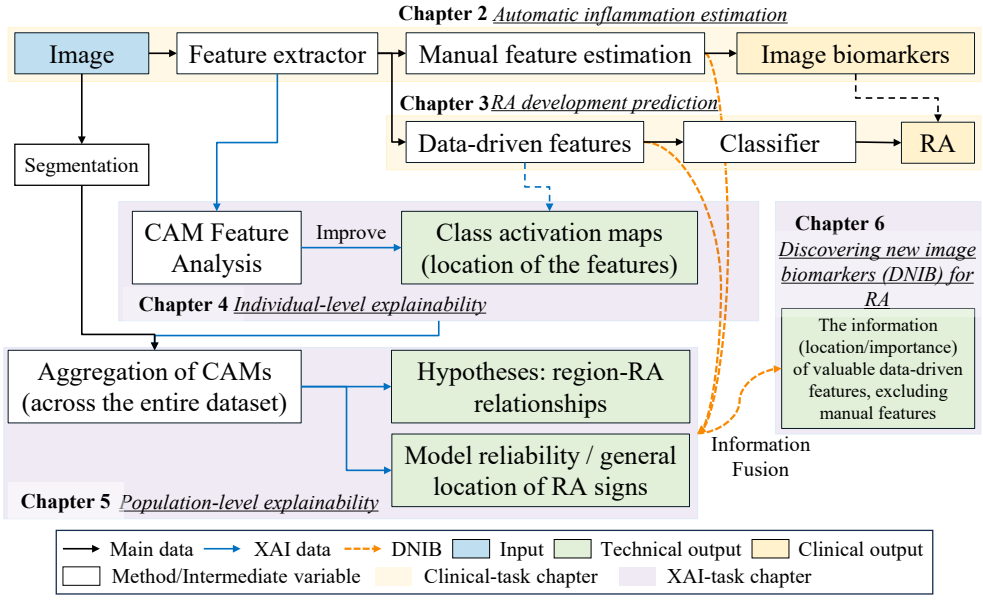


Figure 1.6: A general overview of this thesis, including the tasks described in each chapter and the relationships among chapters. Chapter 2 and 3 focuses on DL models applied to clinical tasks that automatically estimate joint inflammation and predict future RA development, respectively. To validate the reliability of and extract more information from these DL models, Chapter 4 presents a feature analysis framework for class activation maps (CAMs) to fix the intensity scaling problems and to distinguish between feature contributions, and in Chapter 5 further development is presented of the application of CAMs to obtain population-level explainability, based on segmentation. Based on these chapters, Chapter 6 provides a method of discovering new image biomarkers (DNIB) by – aggregating (Chapter 5) properly generated CAMs (Chapter 4) for DL models that predict RA development (Chapter 3) with manually-defined features as intermediate outputs (Chapter 2) – to obtain the importance of all regions with segmentation, excluding the regions of manually-defined features.

This thesis describes a study that aims at developing methods in the detection of early inflammatory signs in RA and predict RA using explainable DL based on MRI (see Fig. 1.6). This study consists of five main components, answering following questions:

- Can DL perform automatic inflammation estimation as accurately as manual visual scoring? (Chapter 2)

- Can DL predict future RA development as accurate as human experts using image biomarkers? (Chapter 3)

- Can these DL models be made explainable at an individual level using CAMs? (Chapter 4)

- Can these CAMs be aggregated so that conclusions can be made at a population-level? (Chapter 5)

- Do DL models look at different features to predict RA, compared to human experts? (Chapter 6)

To answer these questions, several approaches were proposed and developed in these chapters: Chapter 2 and Chapter 3 focuses on the development of DL models for clinical tasks; in Chapter 4 and 5 two different aspects of the CAM approaches are introduced, to improve the explainability of DL models; Through the methods in these chapters, Chapter 6 provides a DL-based method of highlighting and analyzing new image biomarkers other than manually-defined biomarkers.

Chapter 2 introduces automatic DL-based MRI analysis of inflammatory signs in RA (ADMIRA), which aims at automatizing the tedious visual scoring of inflammatory signs in RA. ADMIRA uses pre- and post-processing alongside DL models to estimate inflammation scores given by experts following the RAMRIS system, using MRI scans of 2254 subjects across four study populations (1226, 616, 177 and 236, respectively). The MRI scans were divided into training, monitoring, testing, and validation sets to ensure robust performance evaluation. Similarly, the method in Chapter 4 was applied to validate the DL model's reliability, illustrating its inference process. This study proves the ability of DL models to learn these manually-defined image features (inflammatory biomarkers) and their quantitative values.

Chapter 3 describes a DL framework to predict RA from arthralgia (potential symptom of RA), distinguish RA from other arthritides, and distinguish these arthritides and arthralgia from healthy controls. In Chapter 3, we propose (1) a new pre-processing method to select the most informative slices from 3D MRI scans with irregular image sizes, minimizing the influence of background noise and influence from irregular sizes of anatomical structures; (2) a multi-input DL model architecture to combine and fuse the information from different anatomical sites and views; and (3) a consistency-based loss function to help the DL model focus on RA-related information. We evaluated the framework on a hold-out test set randomly selected from the whole dataset for each task, using five-fold cross-validation and repeat ten times with different random seed to avoid coincidence. Furthermore, we applied an improved version of class activation mapping (CAM) to check whether DL models are making predictions reasonably.

Chapter 4 presents a method for proper intensity scaling to creating comparable CAMs and feature distinction to determining the contribution of each extracted features, which could solve fundamental problems in interpreting CAMs and provide more insight into the trained DL models and datasets. We proposed a framework to statistically analyze DL-extracted features at a population level, determining feature contributions for global intensity scaling and within-layer feature distinction. The global intensity scale standardizes CAMs, achieving high correlation coefficients (R) with model outputs. Within-layer feature distinction identifies overfitting, confounding factors, outliers, redundancies, and principal features. The method was evaluated on eight datasets with different modalities, Pearson's correlation coefficient (R)s between CAMs and outputs were improved by 10.7 – 64.2% after applying the method, achieving near 100% consistencies with models' outputs. This method has become the basis for all DL models' interpretation in this thesis.

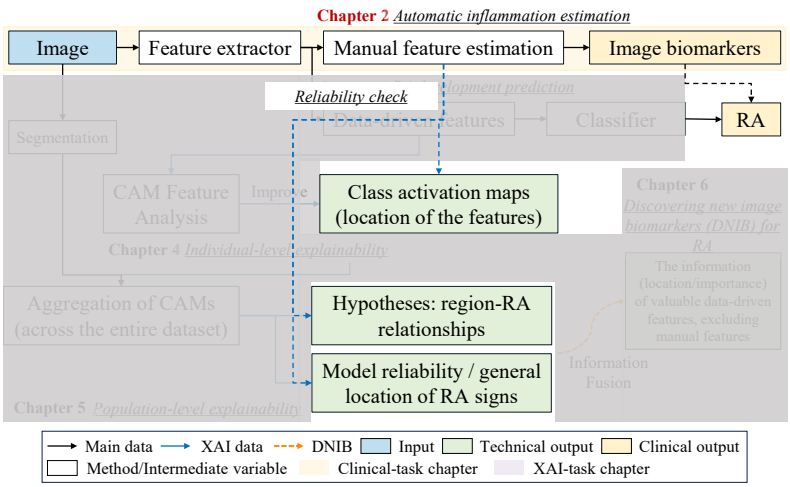
Chapter 5 presents a segmentation-based approach to aggregate the information in CAMs generated through Chapter 4, enabling a quantitative evaluation on how well DL models align with expert knowledge, and allowing to generate fully data-driven hypotheses on the relationship between particular regions or image patterns and clinical outcomes (e.g. particularly the occurrence of TSY or the development of RA in two years). This approach includes a segmentation for locating the image patterns and regions, the CAM generated through the method in Chapter 4. To validate this approach, a series of datasets were simulated based on Bayes' Theorem to have controllable posterior probabilities of outputs given particular image patterns as the ground truths for evaluation. Subsequently, the method was applied to two other medical datasets including our RA prediction task to compare with clinical findings.

In **Chapter 6** we proposed a framework that aims at locating new image biomarkers other than manually-defined inflammatory signs. By combining the fully data-driven route in Chapter 3 with the work in Chapter 2, in this chapter we investigated a new potential of DL in discovering new inflammatory signs. By decoupling manually-designed image features (Chapter 2) from fully data-driven features (Chapter 3) – the residual features after excluding manually-designed features (Chapter 6) could indicate the new inflammatory signs of RA. Through the CAMs from Chapter 4 and aggregation method in Chapter 5, this chapter proposes a naive yet feasible framework to locate these residual features that could indicate early RA.

Chapter 7 and **Chapter 8** summarize the overall findings, implications and future work of this thesis in English and Dutch, respectively.

2

Automatic joint inflammation estimation based on regression neural networks



This chapter was adapted from:

Yanli Li, Dennis A. Ton, Denis P. Shamonin, Monique Reijnierse, Annette H.M. van der Helm-van Mil and Berend C. Stoel. "Automatic joint inflammation estimation from hand and forefoot MRI based on regression neural networks." (*submitted*)

Abstract

Background: Quantitative assessment of inflammation from hand and forefoot MRI scans is crucial for evaluating the severity, progression, and treatment response in inflammatory disease like rheumatoid arthritis (RA). Traditionally, this relies on visual evaluation of signs like bone marrow edema (BME), tenosynovitis, and synovitis, which is time-consuming, subjective, and prone to inherent inter/intra-reader variability.
Purpose: This study aims at an automatic DL-based MRI analysis of inflammatory signs in RA system for inflammation assessment to facilitate related diagnoses and studies.
Methods: We developed an **A**utomatic **D**L-based **M**RI analysis of **I**nflammatory signs in **RA** (ADMIRA) system for inflammation assessment, using pre- and post-processing alongside DL models to estimate inflammation scores from fat saturated, contrast-enhanced T1-weighted MRI scans of 2254 subjects across four study populations. These MRI scans include three different anatomical sites, wrist, metacarpophalangeal (MCP) and metatarsophalangeal (MTP) joints, as the objects for inflammation assessment. The scans were divided into training, monitoring, testing and validation sets to ensure robust performance evaluation, using Pearson's correlation coefficients and Intra-class correlation coefficients. A revised class activation mapping (CAM) algorithm was used to validate the DL model's reliability, illustrating its inference process.
Results: The system achieved mean R/ICCs of nearly 0.9 for synovitis and tenosynovitis on test sets and 0.8 on the validation set, with slightly lower scores for BME (0.8 and 0.7, respectively). This system presents a performance close to human experts on the same datasets. Meanwhile, the visualization results indicate the DL models have a inference process consistent with expert knowledge.
Conclusions: Results show that ADMIRA provides accurate, expert-level inflammation estimation, particularly for synovitis and tenosynovitis, offering a fast, reliable alternative to manual methods for RA monitoring and analysis. We expect that this automatic method could help to reduce labor costs and improve the efficiency of diagnosis in the future.
Keywords: MRI, rheumatoid arthritis, wrist, metatarsophalangeal, metacarpophalangeal, inflammation assessment, deep learning.

2.1 Introduction

Detecting inflammation in the joints of hands and forefeet is essential for recognizing patients at risk of developing inflammatory diseases and for improving patient outcomes. For instance, it helps diagnose rheumatoid arthritis (RA) at an early stage, especially in combination with serological markers [60, 61, 62] . Early diagnoses allow timely treatment, which can improve long-term patient outcomes with higher chances of sustained remission without drugs and improved quality of life [63, 64].

Currently, MRI is the most sensitive imaging method for detecting joint inflammation [24]. In a research setting, MRI scans are assessed semi-quantitatively by visual scoring systems such as the Rheumatoid Arthritis Magnetic Resonance Imaging (MRI) Scoring (RAMRIS) system[31]. Using such scoring systems, trained readers examine multiple tendons or tendon groups, bones and synovia, quantifying the thickness of peritendinous effusion or synovial proliferation and the volume of bone marrow edema. Through this process, such scoring systems provides (semi-)quantitative measurements of multiple image biomarkers, including those describing inflammation severity. However, this is a laborious and difficult task that requires significant time investment, is prone to inter-reader variability and demands rigorous training to ensure accuracy.

To obtain fast and accurate inflammation measurements, an automatic inflammation estimation method could help alleviate labor and time costs, avoiding inter-/intra-reader disagreement, and ultimately improve the diagnosis, monitoring and prediction of inflammatory diseases. However, the automatic quantification of the inflammation in MRI scans is challenging and under-investigated. An accurate, fast and automatic inflammation estimation system that simulates the principles of visual scoring requires not only precise annotations or segmentation of regions of interest (ROIs) on all related anatomical structures in the hands and feet, but also algorithms to identify relevant high-intensity regions indicating of inflammation. In previous studies, Chand et al [35] proposed a segmentation-quantification method to quantify synovitis in MRI scans of 38 patients, following a process of segmenting ROIs and then quantifying the high-intensity pixels in these ROIs. Similarly, Aizenberg et al [36, 25] developed an automatic framework to quantify tenosynovitis and bone marrow edema (BME) in wrist MRI scans of 563 patients, based on atlas-based region annotations and intensity measurements in those anatomical structures. Subsequently, with advances in deep learning (DL), Shamonin et al [37] developed a different framework using DL-based segmentation to obtain annotations for tendons, bones and other anatomical structures. Combined with a fixed threshold, the method measures the volume of high-intensity pixels in the regions of interest, in order to quantify (teno-)synovitis and BME on wrist MRI scans of 1225 patients. Yiu et al [32] developed

a similar segmentation-quantification framework on MRI scans of 80 RA patients, segmenting anatomical structures with DL models and quantifying high-intensity pixels. To mitigate the dependency on accurate annotations of anatomical structures, Mao et al [33] proposed a method for estimating synovitis in MRI scans of 47 RA patients using ROIs instead of annotations. In this study, a classification DL model with different classes that represents different inflammation severity was trained on a small group of manually annotated regions of interest (ROIs) in MRI scans to quantify synovitis and conduct the inference process using ROIs, generated by an unsupervised image segmentation preprocessing rather than accurate annotations or the manual selected ROIs in [34].

However, these methods require accurate annotations or segmentation of ROIs, which are time-consuming processes and can lead to propagation of errors. Furthermore, when extending the system for analyzing MRI scans of other anatomical region (such as scans at the level of the metacarpophalangeal joints in the hand or metatarsophalangeal joints in the feet), it takes considerably more time to obtain accurate ground truth annotations than ground truth visual scores.

A solution to this limitation is the development of an end-to-end automatic framework that directly estimates inflammation scores using DL models without requiring annotations. A framework of mimicking the visual scoring process can be trained on a small group of MRI scans and then applied to other MRI scans without the need for ground truth annotations.

This type of DL-based automatic framework, which quantifies structural pathological symptoms or image biomarkers, have already been widely investigated in general medical image analysis. For semi-quantification tasks that aim at discrete grades or categories, the methods were usually designed to apply classification instead of regression, such as Gleason scoring of prostate cancer in histopathology images [38, 39, 44], grading of ulcerative colitis in endoscopic images [41], diabetic retinopathy grading in eye fundus images [46] and knee osteoarthritis severity classification in MRI scans [43]. For quantification tasks that output continuous scores, regression neural networks are preferred, such as ventricle function indices estimation [42] and arthritis activity scoring based on ultrasound [45], bone erosion scoring based on X-rays [40], Agatston scoring [65], coronary calcium scoring [66, 67] and systemic sclerosis scores [68] based on CT scans. In terms of applications on MRI scans, studies such as grading of abnormalities [69] and osteoarthritis severity grading in knee MRI [70] provided references in building DL-based frameworks for scoring image biomarkers.

Since these studies were not designed for estimating joint inflammation, it is difficult to transfer these ideas or models to our domain. For instance, their model architecture may require specific input shape or image features, which are not feasible in our study. The closest study to our work is a DL-based classification for erosion,

synovitis and osteitis in hand MRI scans [71], which follows a two-step process including semi-automatic localization for regions of interests (ROIs) and DL-based estimation based on ROIs. The semi-automatic localization for ROIs requires human experts to put 3D landmarks corresponds to the 3D centers of anatomical ROIs, slowing down the process and introducing extra labor and time investment. Meanwhile, the DL-model based on ROIs was pretrained through a 3D video classification task and limited to 3D MRI scans with high resolution on all axes, which are not always available . DL models for 3D MRI scans requires considerable amount of data for thorough training and more hardware memory for development and implementation, and will not be useful and generalizable if the cost is even higher than manual visual scoring.

Furthermore, some tasks require more than one MRI scan as input, i.e. multi-view or multi-site inputs (e.g. in systemic diseases with symptoms in different anatomical regions as RA). For these tasks, an ideal method should be capable of coping with memory constraints and fusing the information from multiple inputs. In the specific case of our study, it is important that information from coronal and transversal views are fused, in much the same way human experts do, while performing visual scoring. Both coronal and transversal views are 3D images, but with the highest resolution in either the coronal or transversal plane, respectively.

To overcome these challenges, we propose a DL-based framework that includes automatic preprocessing to select the most informative slices and a DL model to extract and integrate information from both views. This automatic framework utilizes the intensity difference between inflamed and normal tissues to enhance the information density, reduce noise and fuse the information from coronal and transversal scans using cross-attention blocks in the DL architecture. Through learning from the visual scoring system RAMRIS that assess the joint inflammation for RA, this automatic DL-based MRI analysis of inflammatory signs in RA (ADMIRA) system is expected to assess joint inflammation based on hand and forefeet MRIs.

The layout of this paper is as follows. First, we introduce our MRI materials and define the task. Subsequently, we clarify the pre-processing and the architecture of the DL model. We then present experiments evaluating the proposed method for scoring synovitis, tenosynovitis and BME in MRI scans of the wrists, metacarpophalangeal (MCP) and metatarsophalangeal (MTP) joints, collected from 2279 subjects. Finally, we discuss and summarize the limitations and advantages of the proposed methods in the concluding chapters.

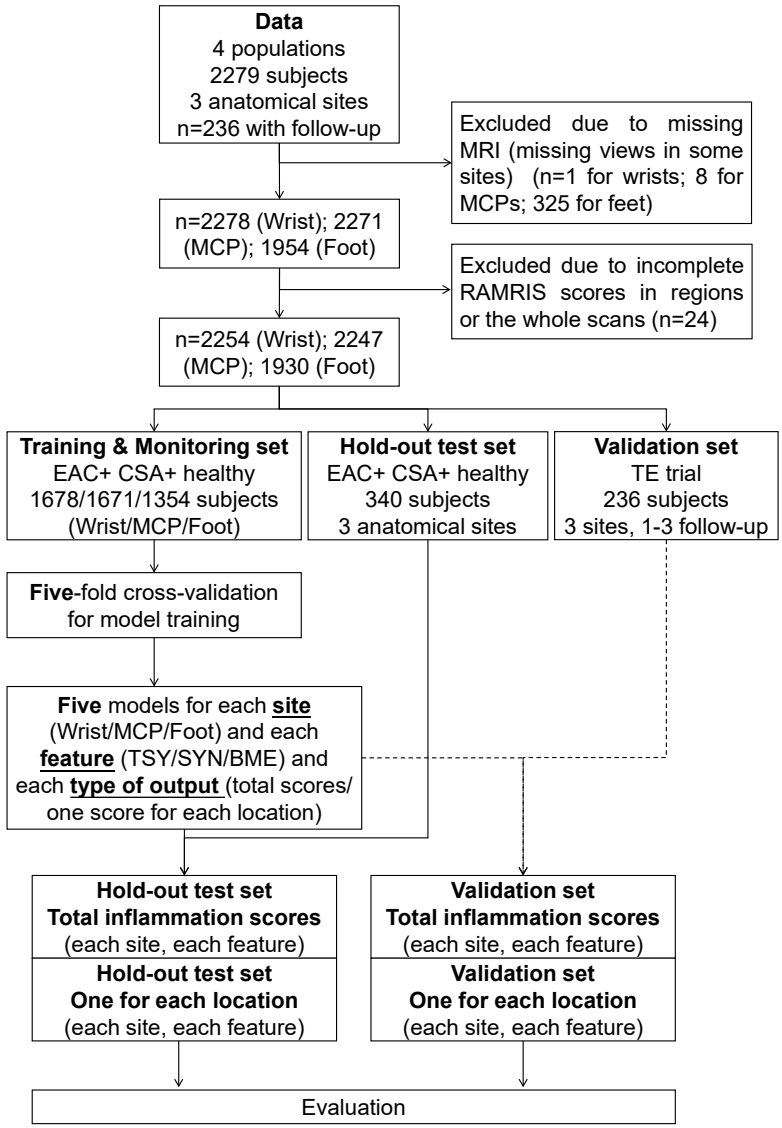


Figure 2.1: Overview of patient selection within the dataset and the data flow during model training and evaluation.

Table 2.1: Characteristics of subjects

Characteristic	EAC	CSA	Healthy controls	TREAT EARLIER trial
No. of patients	1226	616	177	236
Mean age (y) (first time point for longitudinal data) ± SD	56.5 ± 15.6	43.2 ± 12.6	49.8 ± 15.8	46.7 ± 11.9
Female (%)	698 (56.9)	473 (76.8)	136 (70.5)	154 (65.3)
Total RAMRIS inflammation score, median [IQR]	11 [4.5 – 21.5]	2 [1 – 5]	2 [0.5 – 4.5]	4.5 [3 – 8]
Synovitis score, median [IQR]	4.5 [1.5 – 8]	1 [0 – 2]	0.5 [0 – 2.5]	2 [1 – 4]
Tenosynovitis score, median [IQR]	2.5 [0 – 6]	0 [0 – 1]	0 [0 – 0]	1 [0 – 2.5]
Osteitis score, median [IQR]	3 [1 – 6.5]	0.5 [0 – 1.5]	1 [0 – 2]	1 [0 – 2]

Table 2.2: Technical parameters of the hand and forefoot MRI scans.

MRI Parameter	Transversal (axial) scan	Coronal scan
In-plane matrix	320x192	364x224
Repetition time (ms)	570	650
Echo time (ms)	7	17
Echo train length	2	2
Slice thickness (mm)	3	2
Slice gap (mm)	0.3	0.2
Fat saturation	Frequency-selective fat saturation applied	
Scanner	ONI MSK Extreme 1.5T extremity MR scanner (GE Healthcare, Waukesha, WI, USA)	
Coil	100 mm coil	
Contrast	intravenous injection of 0.1 mmol/kg Gd-chelate (gadoteric acid, Guerbet, Paris, France)	
Other	No acceleration, receive-only coils with 4 channels	

2.2 Methods

2.2.1 Data

To develop and assess the ADMIRA system, we utilized a database with four populations with different severity of joint inflammation and containing MRI scans of three anatomical regions: wrists, metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints from a period of more than ten years. In total, 2279 subjects (patient characteristics in Table 2.1) were collected for training the DL models and validating the whole method. Informed consent was given by all patients, referring to the LUMC protocol reference numbers: B19.008 and P11.210. The database consisted of the following four populations: 1247 subjects with early onset arthritis (EAC), 620 patients with clinically suspected arthralgia (CSA), 177 healthy controls and 236 CSA patients with longitudinal MRI scans from the TREAT EARLIER (TE) trial (not included in the CSA group).

As presented in Table 2.2, subjects were scanned with an ONI MSK Extreme 1.5T extremity MRI scanner (GE Healthcare, Waukesha, WI, USA) with a 100 mm coil. After intravenous injection of 0.1 mmol/kg Gd-chelate (gadoteric acid, Guerbet, Paris, France), a T1-weighted fast spin-echo sequence with frequency-selective fat saturation was obtained in the axial plane with a repetition time of 570 ms, echo time of 7 ms,

acquisition matrix 320×192, echo train length 2, slice thickness of 3 mm, and a slice gap of 0.3 mm. Axial images were reconstructed with an image size of $512 \times 512 \times 20 \pm 5$ voxels (voxel size: $0.195 \times 0.195 \times 3.0$ mm). Coronal images were reconstructed with an image size of $512 \times 512 \times 20 \pm 5$ voxels (voxel size: $0.195 \times 0.195 \times 2.2$ mm).

Then, using the visual scoring system based on RAMRIS [22], for each anatomical region, images were independently scored by two trained readers who were blinded to clinical data. For tenosynovitis (TSY) and synovitis (SYN), the readers provided a grade on a scale of 0 to 3, based on the estimated maximum width of peri-tendinous effusion and synovial effusion/proliferation (for TSY and SYN, respectively) with contrast enhancement [22], as follows: grade 0, normal; grade 1, ≤ 2 mm; grade 2, > 2 mm and ≤ 5 mm; grade 3, > 5 mm. The scoring region was bounded proximally by the distal radius/ulna and distally by the hook of the hamate. For bone marrow edema (BME), the readers provided a grade also on a 0 – 3 scale based on the estimated fraction of affected bone volume: 0, no BME; 1, 1 – 33% of bone edematous; 2, 34 – 66%; 3, 67 – 100%. In the following, the mean scores of the two readers are regarded the ground truth and reference for comparison. Based on the above ground truth scores, the inflammation score of each relevant anatomical structure was obtained and served as the training target. However, since it is extremely rare for most anatomical structures to be affected by inflammation simultaneously, this results in a long-tail distribution of the inflammation scores for individual anatomical structures. Consequently, most evaluation metrics are biased and inaccurate, and therefore the evaluation was based on the total inflammation severity - the sum of these scores.

Twenty-five subjects (21 EAC patients, 3 CSA patients and 1 TE trial patient, respectively) were excluded from this study, due to incomplete visual scores or missing MRIs. Subjects' characteristics, technical information (after selection) and dataflow are shown in Table 2.1, 2.2 and Figure 2.1, respectively.

The whole dataset was then split on a subject level, into a training-monitoring set (1684 wrist scans/1677 MCP scans/1362 MTP scans from EAC, CSA and healthy controls), a hold-out test set (340 scans of all anatomical regions from EAC, CSA and healthy controls) and an independent validation set (TE trial); details can be found in 2.1. Under this split, the training-monitoring set also included subjects with missing MRI scans on some of the anatomical regions to maximize the amount of information. The hold-out test set and independent validation set included the patients from whom all anatomical regions were available to fairly compare the performance of models among each region.

To thoroughly evaluate the proposed method and adapt to different uses of the automatic inflammation assessment, the proposed method has two different routes for each input in order to serve different purposes – (1) $Route1$ serves for the situation that a total inflammation score for each inflammatory sign (TSY/SYN/BME) in the

entire anatomical region is needed; (2) *Route2* provides extra information on each relevant anatomical structure, it will firstly estimate one score for each inflammatory sign on each anatomical structure (e.g. tenosynovitis severity around the extensor carpi ulnaris tendon) in the anatomical region (e.g the wrist), and then sum these estimations to calculate the total inflammation score (e.g. tenosynovitis score in the wrist).

To test the robustness of the training process, we applied five-fold cross-validation on the training and monitoring sets to obtain five sets of weights (for each anatomical structure, each inflammatory sign) trained with different data in training sets for the same model architecture. Then the models with the different weights were applied to the hold-out test set and the independent validation set. The robustness of the training process, as the influence of training results on the performance on unseen data, can therefore be checked through the standard deviations in the performance on the evaluation set, using these models with different sets of weights.

2.2.2 Preprocessing

The preprocessing in this ADMIRA system consisted of background removal and slice selection, both were applied to improve the information density of inflammatory signs in the 3D MRI scans. The background removal followed a similar idea in our previous work in rheumatoid arthritis prediction based on MRI scans [72], excluding the background which may contain some high-intensity MRI artefacts or textures. During this background removal, images were (1) thresholded at 10% of the maximum intensity value (optimized manually on a small subgroup of this dataset) to obtain initial foreground masks; (2) then these foreground masks were processed through morphological opening and closing operations [73, 74] to include most of the relevant structures on the borders of the images; (3) subsequently, these foreground masks were applied to the original images to filter out the background; (4) finally, the filtered images were normalized individually and slice-by-slice to a range between 0 and 1, with a 95% clipping to avoid over-normalization caused by the extremely high values from inflamed regions.

The slice selection was applied after the background removal to the filtered images to minimize the influence of background artefacts. For each slice in the coronal scans, the standard deviations of the intensities were calculated, and subsequently seven consecutive slices were selected with the highest standard deviations. For the transversal scans, the central seven slices were selected as they typically fell into the scoring areas in manual scoring systems.

Table 2.3: Configuration of model architecture and training

Modules & Hyperparameters	Configs.
3D Conv Block (1th block)	Conv: depth: 2, kernel size: [3×3×1], number of kernels: 32, stride: 1, group: 2 Activations: ReLU
3D Conv Block (2-4th block)	Conv: depth: 2, kernel size: [3×3×1], number of kernels: 64, stride: 1 Pool: pooling size: [3×3×1], max pool, stride: 2, group: 2 Activations: ReLU
3D Conv Block (5-6th block)	Conv: depth: 2, kernel size: [3×3×1], number of kernels: 96 (5th block) and 160 (6th block), stride: 1, group: 2 Pool: pooling size: [3×3×1], max pool, stride: 2 Activations: ReLU
Cross-attention block (7th block)	In and out channel: 160 Inner dense layer channel: 640 Window size: [2×2] Head: 4 Dimension of head: 8 Dropout rate: 0.3
Adaptive average pooling	Output size: [1×1×1]
Dense layer	Layer 1: [640Œ2 (views)Œ1×1×1, number of scores]
Optimizer	AdamW
Learning rate	4e-5
Weight decay	1e-2
Optimizer momentum	0.9, 0.999
Training epochs	120
Gradient clip	None

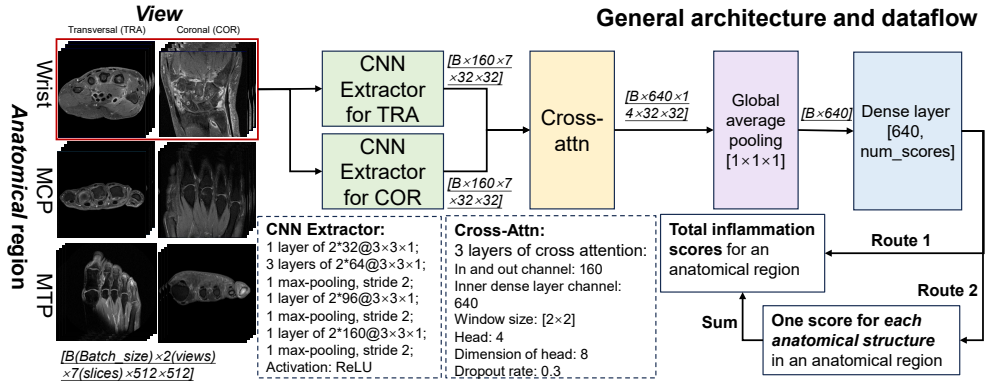


Figure 2.2: The model architecture and data flow.

2.2.3 Model architecture and training

To fuse the information from two views (transversal and coronal), as experts intrinsically do during visual scoring, we applied cross-attention modules [75, 76]. The architecture (see in Fig. 2.2) starts with two separate convolutional neural networks (CNNs) that extract features from the seven slices from the separate transversal and coronal scans. Subsequently, we applied cross-attention modules to exchange the information extracted from the two views. At the end of the cross-attention modules, we used adaptive average pooling and dense layers to obtain the inflammation estimations.

3D images with two views (transversal and coronal) of size [2(view), 512, 512, 13 ± 7] voxels were preprocessed to [2, 512, 512, 7] as input during training. The models were trained to output the inflammation scores on each anatomical region by estimating manual scores, and quantifying the inflamed joints.

As introduced above, two routes were defined in ADMIRA system for different clinical uses: (1) for the situation that the total inflammation assessment is needed, $Route1$ provides a direct estimation for the total inflammation scores (e.g. tenosynovitis for wrists); (2) for the situation that requires an inflammation severity score of each anatomical structure in the anatomical regions (Wrists/MCPs/MTPs), a model is trained to firstly output an estimation for each anatomical structure (e.g. tenosynovitis severity around the extensor carpi ulnaris tendon) and then output the estimation for the entire anatomical region in $Route2$. Therefore, in total 18 models ($3regions \times 3inflammatorysigns \times 2typesofoutputs$) were trained from scratch, using Kaiming initialization [77] on the training-monitoring set, yielding a region-wise estimation (e.g. total MCP tenosynovitis score) or a structure-specific estimation (e.g. tenosynovitis score around the extensor carpi ulnaris tendon). The details of the configurations can be found in 2.3.

2.2.4 Testing, validation and statistical analysis

Images in the test and independent validation sets were preprocessed through the same pre-processing methods as during training, and then used as input for the trained models to generate the scores. For each anatomical region and each inflammatory sign, two models with ten different weights (five for each type of outputs) from the training phase was applied to obtain the final scores. The correlation coefficients were calculated based on the total inflammation scores to have a large range of score distribution and minimize the influence of the long-tail distribution, as the visual scores of each anatomical structure in anatomical regions are integers between 0 and 3 and the vast majority of them are zero. Furthermore, in pre-experiments, the mean squared losses between DL outputs and ground truths presented low consistency with

the correlation coefficients calculated separately for each anatomical structure during training, leading to low training efficiency and poor performance. Therefore, we chose to evaluate based on based on the sum of all structures in each region, for both types of output that have one score for total inflammation and the type of outputs that have one score for each anatomical structure in a specific anatomical region. Consequently, the score range in the calculation has changed from between 0 and 3 to between 0 and $3x$, where x represents the number of defined anatomical structures.

The performance of the proposed method was validated against the performance of human experts, by comparing the agreement between the models' outputs and ground truths with the agreement between readers. This inter-reader agreement then serves as an upper limit for the method's performance, since the ground truth includes the inter-reader variability. We choose Pearson's correlation coefficient R and intraclass correlation coefficients (ICC) as metrics of agreement, as the visual scores are ordinal data. The definitions of these metrics are as below. The $correlation$ between two variables $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ (the scores given by readers or models) is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

where (x_i, y_i) are the individual sample points, $(\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i)$ are the sample means of X and Y, n is the number of paired samples (i.e. number of subjects). The ICC used in this study, $ICC(2,1)$, between two variables $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ (the scores given by readers or models) can be defined through the following process:

$$\mathrm{ICC}(2,1) = \frac{2S^2_{\text{between}} - S^2_{\text{error}}}{2S^2_{\text{between}} + S^2_{\text{error}} + \frac{2}{n}(S^2_{\text{rater}} - S^2_{\text{error}})} \tag{2.2}$$

where $S^2_{\text{between}} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{z}_i - \bar{z})^2$ is *Between-subject variance* that describes variance between subjects, $S^2_{\text{error}} = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \bar{z}_i)^2 + (y_i - \bar{z}_i)^2\right]$ is the *Residual (within-subject) variance* that describes variance between readers, $S^2_{\text{rater}} = (\bar{x} - \bar{z})^2 + (\bar{y} - \bar{z})^2$ is the *Mean difference between raters* that describes general bias, n is the number of subjects. $ICC(2,1)$ was applied instead of other ICCs because we aimed to evaluate whether $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ were consistently rated. Meanwhile, ICC between human readers gives an upper limit for automatic methods trained on the scores.

Furthermore, to prove the effectiveness of the proposed slice selection and model architecture, we compared the performance with some existing methods and the ADMIRA models without the some proposed strategies. As the baseline methods in other studies were not designed for the input and the anatomical objects, we used

the same pre-processing (including the slice selection) in these existing methods as in our models, and used an extra dense layer to fuse the information extracted from two input views.

All experiments were executed on an RTX6000 GPU from Nvidia, with PyTorch 1.12 https://pytorch.org/ on Python 3.9 https://www.python.org/ and SciPy 1.7 https://scipy.org/. These were executed ten times with random seeds, and the results of the models with median performance are presented. Details on the configurations and model architectures, configurations and a simple inference application can be found in the online open-source repository https://github.com/YanliLi27/ADMIRAinfer.

2.2.5 Reliability and explainability

One substantial problem in the deployment of DL models in medical image analysis is their reliability. Although DL models are "black boxes" with large numbers of parameters and difficult to comprehensively ensure their robustness, an intuitive way of facilitating their applications is to improve the explainability. To ensure that DL models in our study were quantifying inflammation based on really inflamed regions instead of artifacts or some other confounders, we applied a revised version [57] of the class activation mapping (CAM) technique [59, 58, 55], one of the most common explainability techniques for deep learning, to open the black box. The revised method, called rescaled regression activation mapping [57], generates heatmaps that highlight the most informative regions, which were expected to be the inflamed regions.

2.3 Results

The ADMIRA system obtained mean ICCs of around 0.9 and 0.8 on the test set and validation set, respectively, for all anatomical regions and inflammatory signs. The same numerical results were obtained in terms of Pearson correlations. For each inflammatory sign and anatomical structure, two outputs are given by the two routes: one total score (synovitis/tenosynovitis/BME) for the whole anatomical region (wrists/MCPs/MTPs) and a series of scores for each anatomical structure (e.g., WRTSYI for tenosynovitis in wrists, MCSYN3 for synovitis in MCPs). Therefore, 18 outputs (three inflammatory signs × three anatomical structures × two routes) are presented in the following sections. The results of the proposed inflammation estimation method is firstly compared with the performance with human experts for indicating the general performance, and then compared with the other DL frameworks and models to show the effectiveness of the whole framework. The ADMIRA system takes a median runtime of less than 0.1 second per input under the training soft- and hardware setting, and nearly 1 second per input using normal CPUs.

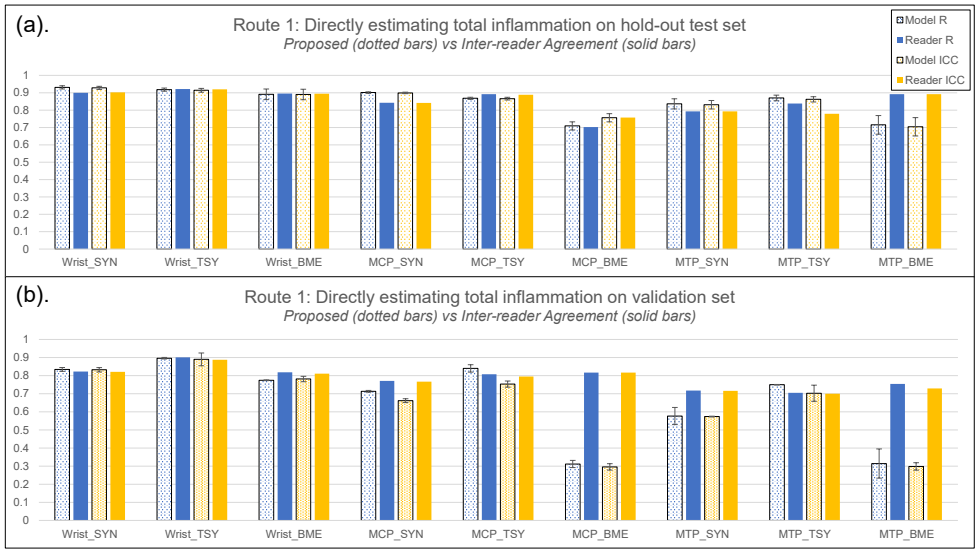


Figure 2.3: The performance of the ADMIRA system on directly estimating total inflammation scores on (a) hold-out test set and (b) independent validation set. The dotted bars represent the performance of the proposed method and the solid bars represent the inter-reader agreement; the blue bars represent the correlation coefficient, R, and the orange bars represent the intra-class correlation coefficients (ICCs). The error bars on the dotted bars, from the models with different weights during the five-fold cross-validation training process on training and monitoring set, represent the training robustness. The results demonstrate that the trained model based on $Route1$ performed close to human experts on assessing synovitis and tenosynovitis, yet worse on assessing bone marrow edema.

2.3.1 Performance of Route 1 and comparison with human experts

Fig. 2.3 presents the general performance of the ADMIRA system on estimating total inflammation scores for the whole anatomical region, on both the hold-out test set (with a similar inflammation distribution as the training and monitoring sets) and independent validation set (with a different inflammation distribution).

As shown in Fig. 2.3 (a), the ADMIRA system received Rs and ICCs around 0.9 on estimating total inflammation scores of synovitis, tenosynovitis and BME in wrists and (teno-)synovitis in MCPs, while receiving Rs and ICCs of nearly 0.8 for BME in MCPs and tenosynovitis in MTPs. By comparing the inter-reader agreement on these total inflammation scores according to the manual scoring system, the ADMIRA system achieved a level close to the expert level on the hold-out test set, especially on estimating tenosynovitis and synovitis.

From Fig. 2.3 (b), compared to the performance on the hold-out test set, the proposed method received Rs and ICCs of around 0.8 for inflammation in wrists, (teno-)synovitis in MCPs and tenosynovitis in MTPs. However, the ICCs on estimating BME for MCPs and MTPs dropped from 0.7 to 0.3 while the Rs remained around 0.8 and 0.7. These results indicate that the inflammation assessment on tenosynovitis and synovitis is promising, while the assessment on BME based on the ADMIRA system needs further investigation before application.

2.3.2 Performance of Route 2 and comparison with human experts

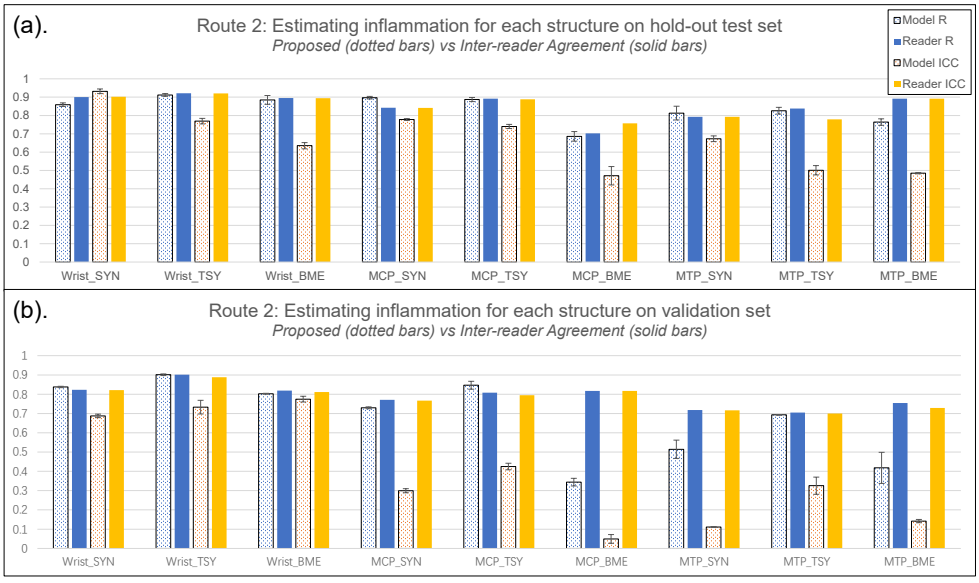


Figure 2.4: The performance of the ADMIRA system on estimating inflammation scores through summing up the estimated inflammation scores of each anatomical structure on (a) hold-out test set and (b) independent validation set. The dotted bars represent the performance of the proposed method and the solid bars represent the inter-reader agreement; the blue bars represent the correlation coefficient, R, and the orange bars represent the intraclass correlation coefficients (ICCs). The error bars on the dotted bars, from the models with different weights during the five-fold cross-validation training process on training and monitoring set, represent the training robustness. The results demonstrate that the trained model based on *Route2* performed close to human experts on assessing inflammation in the wrist, yet worse on assessing MCP and MTP joints if the distribution of inflammation severity is different.

As mentioned, *Route2* firstly generates a score for each anatomical structure in the anatomical region and then sum these scores to obtain the total inflammation scores. Fig. 2.4 presents the general performance of ADIMRA *Route2* on both hold-out test set and independent validation set. The numbers of scores for each inflammation sign

and anatomical region are: 10 for wrist tenosynovitis, 3 for wrist synovitis, 15 for wrist BME, 8 for MCP tenosynovitis, 4 for MCP synovitis, 8 for MCP BME, 10 for MTP tenosynovitis, 5 for MTP synovitis and 10 for MTP BME.

As shown in Fig. 2.4, the performance of the ADMIRA system on estimating inflammation scores through $Route2$ – firstly estimate a score for each anatomical structure and then summing the estimates to obtain the total scores – achieved a promising level, with Rs and ICCs around 0.9 on the hold-out test set and 0.8 on the independent validation set. The ICCs on estimating synovitis and tenosynovitis in the validation set significantly decreased due to a long tail distribution of the inflammation scores. Generally, compared to the performance of $Route1$ that directly estimate total inflammation scores for the entire anatomical region, $Route2$ is not as competitive as $Route1$ in MCPs and MTPs, especially when estimating inflammation scores for MTP and BME. However, $Route2$ provides the estimation for each anatomical structures in the anatomical regions, enabling more clinical uses that requires the inflammation severity for specific structures.

2.3.3 Output distribution of ADMIRA inflammation assessment

Fig. 2.5 present the scatter plots of the ground truths (X-axis) and the estimated scores from the proposed method (Y-axis) of $Route1$ and $Route2$, respectively. In these scatter plots, the green diagonal lines represent the ideal correlation, and the red lines represent the fitted line of the proposed method. Generally, from the perspective of inflammatory signs, the proposed method performed the best on estimating tenosynovitis and worst on estimating BME; regarding anatomical regions, the ADMIRA system performed the best on wrists and worst on MTPs; in terms of datasets, ADMIRA performed better on the hold-out test set than on the independent validation set – probably because the hold-out test set has an inflammation distribution close to the training and monitoring set. The distribution of errors over the different inflammation severities is relatively uniform, except for some outliers in estimating BME.

2.3.4 Comparison with existing methods

Fig. 2.6 presents the comparison of the proposed method and some existing methods. As these baseline methods were not designed for the input and the anatomical objects, we used the same pre-processing (including the slice selection) as we applied to our models and used an extra dense layer to fuse the information extracted from two input views. In addition, Fig. 2.6 presents the results using the same architecture with 2D Convolutional and 2.5D inputs and using random slice selection to show the effectiveness of the proposed preprocessing and 3D inputs.

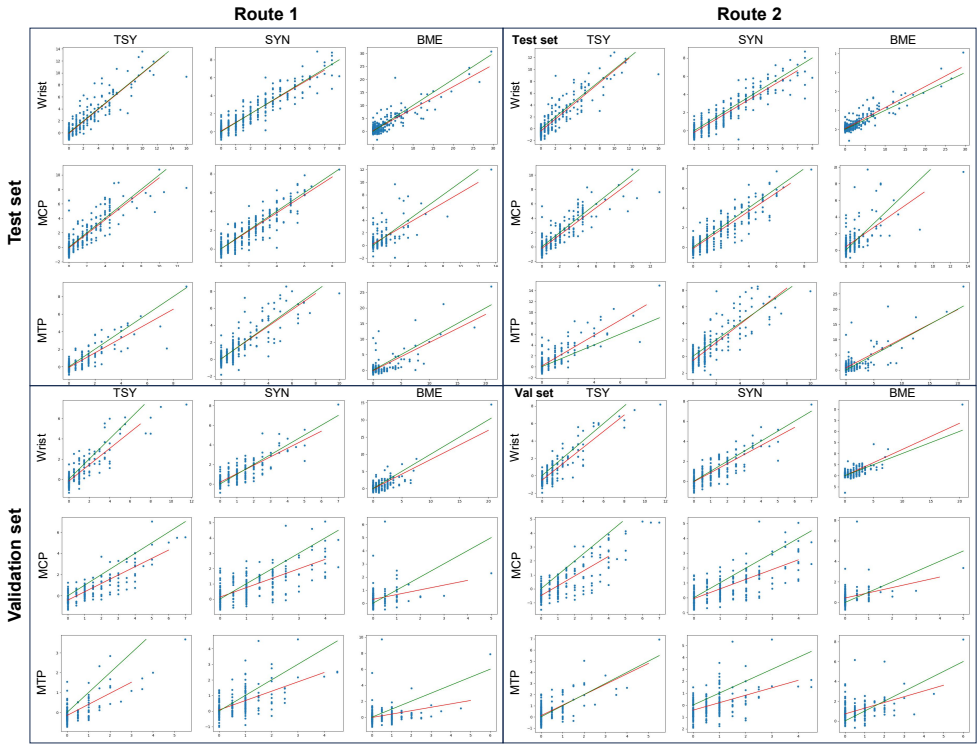


Figure 2.5: Scatter plots of the ground truths (X-axis, the mean scores of two readers) and the estimated scores from the ADMIRA system (Y-axis) when estimating total inflammation scores through $Route1$ and when estimating inflammation through $Route2$.. The green diagonal lines represent the ideal relation and the red lines represent the fitted line of the proposed method.

The results between our model with and without the slice selection demonstrate the effectiveness of the proposed preprocessing. Meanwhile, the comparison with other model architectures indicates the effectiveness of the Transformer-based information fusion. The comparison between the 2D and 3D version of ADMIRA suggest 3D inputs with 3D Conv could preserve more information than 2.5D inputs and 2D Conv.

2.3.5 Explainability through CAM

Figure 2.7 presents some examples of slices in the CAMs generated from the DL models in ADMIRA trained for the estimation of total inflammation score directly ($Route1$). The visual checks of the CAMs suggest that the DL models of ADMIRA followed a principle similar to visual scoring, and to some extent proved the reliability of the proposed method. Moreover, the CAMs also suggest some potential correlation

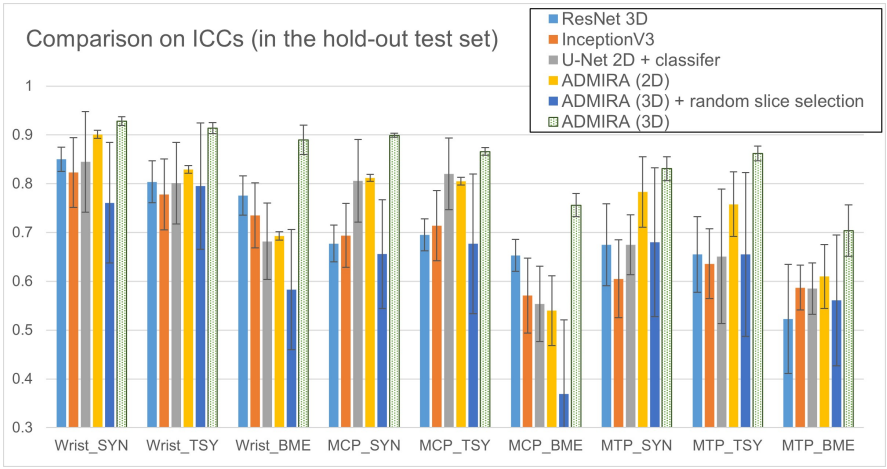


Figure 2.6: Comparison with baseline methods regarding ICCs and total inflammation estimation ($Route1$) on hold-out test set, including ResNet 3D [71], InceptionV3 [38, 41], U-Net encoder with a classifier [78], ADMIRA with 2D convolutional blocks, our 3D model with a random slice selection. Except for the 3D model + random slice selection, all models were fed with seven selected slices and the slice dimension served as a channel for 2D models and as third dimension for 3D models. Complete MRI volumes as input were tested and received lower performance compared to slices as input, and is therefore not presented. Our method with 3D models (fed with selected slices) yielded a significantly greater similarity compared to other methods, as supported by a p-value less than 0.05.

between different inflamed regions, as tenosynovitis regions received higher signal intensities compared to the background, and the intensities of the background should represent no contribution to the evaluation of inflammation.

2.4 Discussion

To our knowledge, this ADMIRA is the first study to develop an end-to-end automatic tool for joint inflammation scoring of rheumatoid arthritis with such a performance close to human experts. The 3D version of ADMIRA system successfully measured joint inflammation in three anatomical regions, achieving high ICCs with the ground truth from visual scoring on fat saturated, contrast-enhanced T1-weighted MRI scans. The automatic estimations of tenosynovitis and synovitis achieved an accuracy close to that of expert assessments, while the challenges in estimating BME require further investigation and validation.

2.4.1 Data distribution

The relatively poor performance in estimating BME may stem from two main factors: (1) the image features of BME in MRIs (such as intensity and boundary) are not as

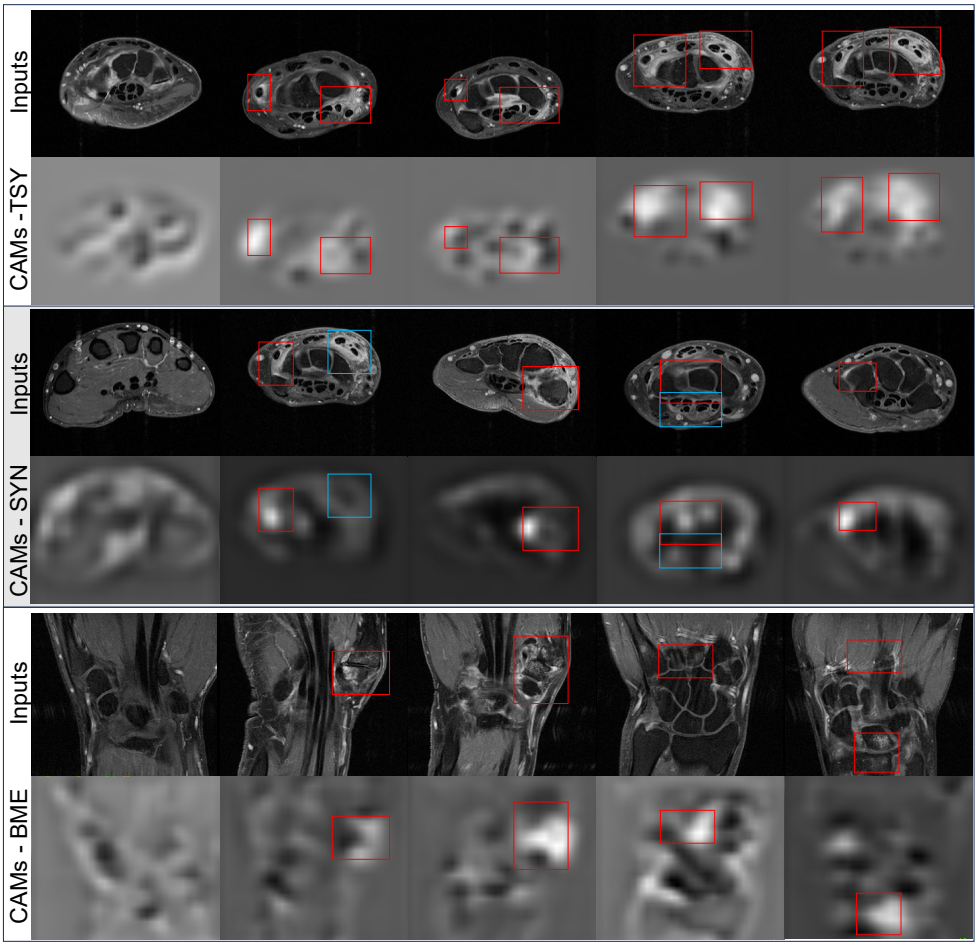


Figure 2.7: Examples of slices from the CAMs of 15 patients generated from the DL models for estimating total inflammation directly ($Route1$). Red boxes in the examples highlight the inflammation regions in the MRI scans that are the target regions for inflammation assessment; blue boxes highlight examples of inflamed regions that belong to another category (e.g. tenosynovitis regions when the CAMs were generated from the DL models for synovitis estimation). CAMs are displayed as separate images instead of the commonly used overlays on the original images, in order to allow for an accurate interpretation by clearly distinguishing class activations from pixel intensities.

clear as synovitis or tenosynovitis, therefore making it difficult for the method to learn consistent patterns of BME; and (2) the estimation of BME faces a considerable imbalance in the distribution of inflammation severity across the whole dataset. Most MRI scans in all the populations studied received a score between 0 and 1 in each anatomical structure and between 0 and 3 in total score, based on RAMRIS. This imbalanced, long-tailed distribution (see Figure 2.8) affected the training of DL models and increased the difficulty of obtaining high ICCs and Rs. Some compensations such as data augmentation and over-sampling were conducted to mitigate this distribution problem. However, inflammation in each anatomical structure is not evenly distributed (some anatomical structures are more susceptible to inflammation) and sometimes correlated with each other, the compensations were not enough to solve the problem. A potential solution to this data imbalance is advanced data augmentation [37] or synthesis [79] that controllably over-sample certain rare cases.

Similar factors also affected the performance of the ADMIRA system on the independent validation set, where the distribution of inflammation severity significantly differed from that of the training, monitoring and hold-out test sets. This discrepancy across the independent validation set derived from the inclusion criteria of the TE trial – TE trial only consists of patients with more subclinical joint inflammation than the healthy controls and the patients included in CSA, and more importantly, the patients with most severe joint inflammation were not included. These different inclusion criteria lead to more patients with only mild joint inflammation and less patients with severe joint inflammation in TE trials, compared to the other study populations, in which EAC patients with severe joint inflammation were included and mitigated the influence of long-tail distribution. Consequently, the visual scores for all kinds of joint inflammation in the independent validation set present a longer-tailed distribution that undermines the performance.

The long tail distribution also explains why ADMIRA performed worse in estimating the scores for each anatomical structure than estimating the total scores directly. The visual scores for each anatomical structure across all scans and all anatomical regions (wrists, MCPs and MTPs) presented severely long-tailed distributions: for each anatomical structure, the score was 0 in the vast majority of scans. The DL models trained to estimate the visual scores for each anatomical structure were therefore facing a similar data imbalance and received moderate performance compared to the models estimating the total scores directly.

2.4.2 Limitations

Besides the consequences of the imbalance in data distribution, the limitations of ADMIRA are of great importance to discuss. The first limitation is that the

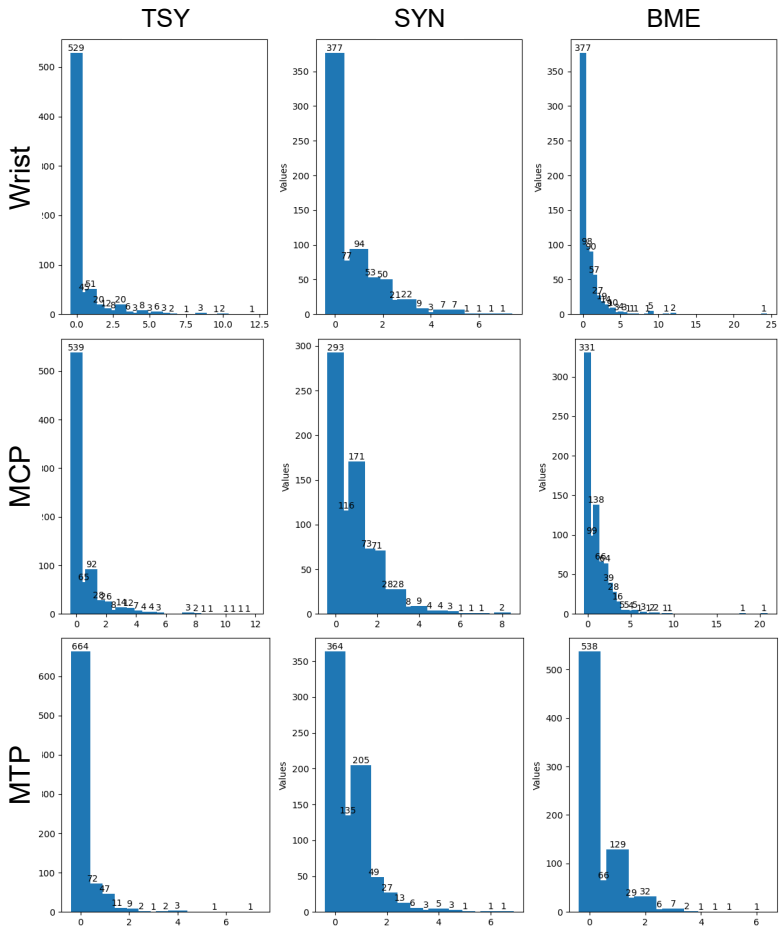


Figure 2.8: The inflammation score distribution of each inflammatory sign in the group of validation set.

method is designed for the described extremity 1.5T MRI protocol with contrast agent administration. While the ideas can be transferred to other MRI sequences or field strengths, some specific adaptations and refinements are required for application to other protocols, and the effectiveness of these ideas cannot be ensured on those datasets without challenges similar to our tasks.

The second limitation is that the performance of DL models may be overestimated due to the potential correlation among inflamed regions, leading to an inflammation estimation on one anatomical structure that originates from a prediction based on another inflamed region instead of the target region.

The third limitation comes from the fundamental basis of the model training, in which visual scores were considered as the ground truths. On the one hand, RAMRIS

scores are the best reference we could access in assessing joint inflammation based on MRI scans; on the other hand, training on these visual scores as golden standards has inherent drawbacks: (1) RAMRIS has a semi-quantitative scale, ignoring subtle differences in scores of less than 1 unit, that could have negative impact on training a model with continuous outputs; (2) Some inflamed tissues are not included in the RAMRIS as it is designed for rheumatoid arthritis, limiting the application range of the proposed method; and (3) The proposed method cannot surpass the performance of RAMRIS, and the quality of ground truths (e.g. inter-reader agreement, see in the appendix) used for training could convey implicit biases and influence the reliability of the proposed method.

The fourth limitation derives from the MRI acquisition protocol. While MRI with contrast agents can provide additional information on active inflammatory sites, these contrast agents could cause concerns such as side effects and are limited under some circumstances. The ability of detecting and assessing inflammation in MRIs without contrast agent is therefore of more significant implications. However, ADMIRA was trained based on fat saturated, contrast-enhanced T1-weighted MRI scans, its ability on MRI scans without contrast agent requires further investigation. Therefore, we are aiming at developing a model based on MR images without contrast enhancement using the current model as a pretrained model in the following projects.

Furthermore, the issue on explainability caused by choosing end-to-end training is a major disadvantage of the proposed DL method, compared to the previous segmentation-quantification approach [11, 37] or landmark-quantification approach [70]. This disadvantage also affects the applications of other methods that are aiming at end-to-end score estimation, as mentioned in introduction. To validate the DL models, we applied the improved regression activation mapping algorithm to generate heatmaps that highlight important regions in MRI scans during the estimation of inflammation severity based on DL models. To some extent, the explainability of the DL models could be verified by comparing the anatomical structure of highlighted regions in the heatmaps and the inflamed regions, proving DL models were estimating inflammation correctly based on the truly inflamed regions in the MRI scans. However, evaluating the reliability by an observer study faces challenges including the experiment design and comprehensive understanding of both in-depth clinical and technical details. For example, as deep learning models can learn from the correlation in inflammation between two regions, the CAMs may also highlight multiple regions out of the ROIs, leading to difficulties in a comprehensive observer study. Considering the importance of the explainability, we are currently working on a comprehensive clinical study in cooperation with clinicians on the reliability, generalizability, efficiency and other perspectives of the proposed method.

2.4.3 The two routes for inflammation estimation

The two routes in this study to obtain the inflammation estimation are designed for different purposes. For clinical uses that requires inflammation assessment for some specific structures in the anatomical regions, $Route2$ provides a structure-wise estimation that can be flexibly selected, summed and then applied to different aims. For the situation that only the total inflammation severity is needed, $Route1$ enables a more accurate estimation, allowing the model to infer and "predict" total inflammation based on any inflammatory signs in the images without precisely estimating inflammation of each structure.

2.4.4 Summary

Despite the above limitations and the need for a comprehensive observer study for explainability, the end-to-end DL models brought not only expert-level accuracy on inflammation estimation, but also a fast calculation in less than 0.1 second (GPU) and 1 second (CPU) for each input and independency on accurate manual annotations or landmarks. These advantages allow the proposed method to be performed on other similar clinical problems and datasets, contributing to reducing the labor and time costs of inflammation assessment.

2.5 Conclusion

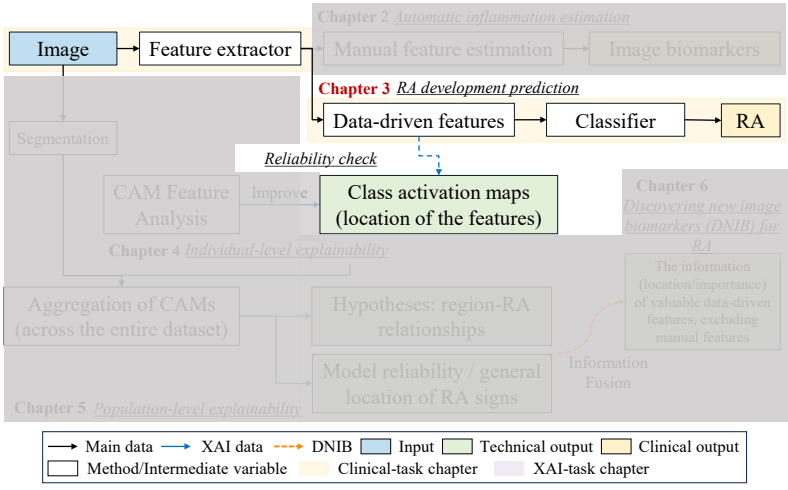
In this study, we proposed a fully automatic system for scoring inflammation from extremity MRI scans called ADMIRA, quantifying the inflammatory signs in (teno-)synovia and bones of the wrists, MCPs and MTPs. The deep learning model learned the meaning of inflammatory areas on fat saturated, contrast-enhanced T1-weighted extremity MRI scans and provided accurate and fast inflammation scoring for wrists, MCPs and MTPs, especially on synovitis and tenosynovitis, based on principles similar to human assessments. We expect that this automatic method could help to reduce labor costs and improve the efficiency of diagnosis in the future.

Acknowledgment

This work is supported by the Netherlands Organization for Scientific Research (NWO, TTW 13329), the ERC (European Research Counsel) starting grant under the European Union's Horizon 2020 research and innovation programm No.714312 and the China Scholarship Council No.202108510012.

3

Rheumatoid arthritis classification and prediction by consistency-based deep learning using extremity MRI scans



This chapter was adapted from:

Y Li, T Hassanzadeh, DP Shamonin, M Reijnierse, AHM van der Helm-van, BC Stoel. "Rheumatoid arthritis classification and prediction by consistency-based deep learning using extremity MRI scans." In *Biomedical Signal Processing and Control*, 2024, 91: 105990.

Abstract

Predicting the development of rheumatoid arthritis (RA) in an early stage through magnetic resonance imaging (MRI) can initiate timely treatment and improve long-term patient outcomes. Although manual prediction is time-consuming and requires expert knowledge, automatic RA prediction has not been fully investigated.We present a consistency-based deep learning framework to classify and predict RA automatically and precisely, including an output-standardized model, customized self-supervised pretraining and a loss function that is based on label consistency between original and augmented inputs. For training and evaluation, we used a database, containing 5945 MRI scans of carpal, metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints, from 2151 subjects obtained over a period of ten years. Four (three classification- and one prediction-) tasks were defined to distinguish two patient groups from healthy controls and RA from other arthritis patients within the recent-onset arthritis group, and predict RA development in a period of two years within the clinically suspect arthralgia group. The proposed method was evaluated with the area under the receiver operating curve (AUROC) on a separate test set, achieving promising mean AUROCs. This proves the existence of early signs of RA in MRI and the potential of a consistency-based deep learning model to detect these early signs and predict RA.

3.1 Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disorder that especially affects joints in wrists, hands, and feet [3]. It can ultimately result in bone erosions and joint deformations and only very early detection and treatment can improve the long-term outcome [7]. Finding early signs, localizing lesions, and predicting potential development into RA can help radiologists and rheumatologists to diagnose and treat RA at an early stage. Therefore, this motivated our study to find early RA-relevant signs through imaging. Magnetic resonance imaging (MRI), which enables the visualization of both anatomical information and inflammatory signs, is the most sensitive imaging method to detect inflamed areas and has become a common imaging modality for RA research. RAMRIS (rheumatoid arthritis magnetic resonance imaging scoring system) [22] is currently the most widely-used imaging biomarker to quantitatively score RA for each anatomical site [25, 80, 81]. To classify and predict early inflammatory signs, RAMRIS assesses bone marrow edema [80], synovitis [82], and tenosynovitis. However, scoring these biomarkers is time-consuming, requires expert training, and depends on prior knowledge and assumptions to detect early signs.

In previous work, automated biomarker quantification methods were proposed and demonstrated a high correlation with expert RAMRIS scores [25]. These pre-defined image features may, however, not be the optimal biomarkers to classify and predict RA. Moreover, certain inflammatory signs may not be relevant to RA, as they also appear in healthy individuals. Therefore, the visual scoring by RAMRIS also compares with healthy controls. This makes it challenging to classify and predict RA through traditional image analysis methods.

Since these tasks typically include labeling or classification, deep learning (DL) methods are highly suitable, without relying too much on prior assumptions or pre-defined imaging biomarkers. Despite the success in other medical imaging labeling tasks [47, 48, 49], DL methods have not been fully investigated yet in the classification and prediction of RA due to the following reasons. Firstly, the time window is narrow for collecting images from arthralgia patients with possible early signs of RA, complicating data acquisition and resulting in a limited dataset size. Consequently, overfitting becomes a severe problem, and the size of the dataset also restricts the performances of large models with massive trainable parameters. Secondly, the variety and complexity of anatomical and pathological structures in hands and feet, and variability in positioning hands and feet, further amplifies the difficulty of the tasks, resulting in insufficient performances of standard models for medical images. Thirdly, artifacts caused by fat suppression errors, movement or aliasing may significantly influence the automatic interpretation of MRI scans. These artifacts may appear more

often in certain time periods, becoming serious confounders while splitting the dataset for evaluation, and worsening the overfitting problem. Finally, there are no publicly accessible datasets for similar subjects or tasks, therefore DL models cannot benefit from transfer learning and well-developed pre-processing methods.

To overcome these challenges and predict RA in an early stage, we propose a so-called consistency-based training framework for a simple deep learning model to pre- and post-process MRI scans, and predict the development of RA in patients with recent-onset arthritis or clinically suspect arthralgia. This consistency-based framework helps the model to utilize the unchanged information and learn from a limited number of samples. Specifically, we first pre-trained the model with a self-supervised reconstruction method, based on a masked autoencoder (MAE) [83], to let the model understand the anatomy of human hands and feet by filling in the masked areas in MRIs from the training set. Meanwhile, we applied an extra contrastive loss function based on augmentation to emphasize that the disease-related information should be invariant to spatial transformations (i.e., the output probabilities or logits should be independent of an object's position or orientation).

Main contributions of this chapter are: (1) This is the first MRI-based early RA prediction framework using deep learning with promising results; (2) A self-supervised reconstruction is applied for pre-training to utilize the anatomical consistency of human hands and feet, thereby replacing Transformers [84] by fully convolutional networks (FCNs) that have far less parameters than the visual learner [83]; (3) A contrastive loss function is defined to accelerate the training process and force the model to focus on unchanged RA information after augmentation.

The layout of this chapter is as follows. First, we introduce our MRI materials and the task definition. Subsequently, the preprocessing, backbone models and the consistency-based deep learning framework are successively explained. Thereafter, we present the overall task performance, general improvements compared to baseline models and ablation studies for input preprocessing, model, pretraining and proposed methods. Finally, the limitations and advantages of the proposed methods are discussed and summarized in the last two chapters.

3.2 Materials

3.2.1 Structure of materials

The models were trained and evaluated based on a database (informed consent given by all patients, LUMC protocol reference number: B19.008 and P11.210) that contained a total of 5945 MRI scans of carpal, metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints, from 2151 subjects obtained over a period of ten years (see Fig. 3.1). This MRI dataset consists of three groups: 1247 patients with recent-onset arthritis, called early arthritis clinic (EAC), 727 arthralgia patients with

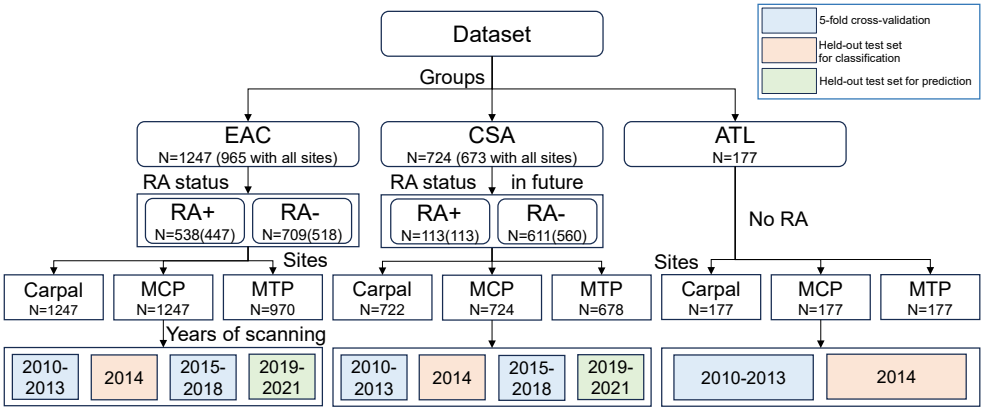


Figure 3.1: Dataset composition, task definition, and evaluation design. The dataset consists of three major groups: patients with recent-onset arthritis (EAC), CSA and healthy controls (ATL). The EAC group was divided into EAC(RA+) and EAC(RA-), while the CSA group was divided into CSA(RA+) and CSA(RA-), where "RA+" / "RA-" indicates the RA status one or two years after the baseline. Each group contains MRI scans collected from the carpal (wrist), metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints. The dataset was collected between 2010 and 2021, while the ATL group was collected over a shorter time period (between 2010 and 2014).

an increased risk of developing RA, called clinical suspect arthralgia (CSA), and 177 healthy controls as atlas (ATL). Study protocols for the EAC cohort (reference number: B19.008, date: 29-may-2009) and CSA/healthy controls(ATL) cohort (reference number: P11.210, date 08-feb-2012) were approved by the local Medical Ethical Committee of the Leiden University Medical Center (LUMC).

The EAC group consists of patients with clinically confirmed arthritis, of which a subgroup was diagnosed with RA within a year, whereas the remainder was diagnosed with other arthritides (non-RA) or undifferentiated arthritis (UA). According to these diagnoses after one year, EAC patients were divided into either RA or non-RA/UA, indicated by EAC(RA+) and EAC(RA-). The classification task was to distinguish these two subgroups. The CSA group was followed over a period of two years in order to establish whether they had developed RA. The CSA group was divided into two groups CSA(RA+) and CSA(RA-), with the task to distinguish these two subgroups, so as to predict the development of RA. The ATL group was collected over a shorter time period. Further details (including patients' characteristics) of the collected dataset can be found in [25].

In each group, the carpal, MCP, and MTP joints were scanned with a 1.5T extremity MRI scanner (GE Healthcare) using a 100-mm coil, with contrast enhancement (T1-

Gd) and frequency-selective fat saturation. For COR scans (3D scans with the highest resolution in the coronal plane), the repetition time was 650 ms, echo time 17 ms, acquisition matrix 364x224, echo train length 2, slice thickness 2mm, and slice gap 0.2 mm. For TRA scans, these parameters are: 570 ms, 7 ms, 320x192), 2, 3 mm, and 0.3 mm, respectively [25]. The scans were reconstructed into [512, 512, 20±5] images, which means the resolution in the Z direction was relatively low, leading to information loss and thus increasing the importance of fusing information from coronal and TRA scans.

3.2.2 Task definition

Table 3.1: Description of the four tasks

Task	Materials
1. Classification into recent-onset arthritis and healthy	EAC, ATL
2. Classification into clinically suspect arthralgia and healthy	CSA, ATL
3. Classification into RA and non-RA/UA, as diagnosed after 1 year	EAC(RA+), EAC(RA-)
4. RA prediction in clinically suspect arthralgia patients	CSA(RA+), CSA(RA-)

To incrementally increase the complexity of training the convolution neural networks (CNNs), we first defined two tasks of making a distinction between two populations (classification task): Task 1 to distinguish recent-onset arthritis from healthy; Task 2 to distinguish CSA from healthy. Task 3 was to distinguish RA from other arthritides and undifferentiated arthritis, as diagnosed after one year, within the EAC group; and Task 4 was to predict future RA development from baseline MRI scans within the CSA group. For pre-training in Tasks 3 and 4, we used the trained encoders from Task 1 and 2, respectively. (see Table. 3.1).

3.3 Methods

3.3.1 Overall workflow

Fig. 3.2 presents the overall workflow and basic information of the proposed methods for training the CNNs for the four tasks. The MRI scans from different anatomical sites were processed by a unified process. The process begins with preprocessing to standardize the output, removing background noise and artifacts, resizing the anatomical structures in the images into a fixed size, slice-by-slice intensity normalization and selecting the central slices for axial scans to increase the information density. This is followed by a simple model pre-trained through self-supervised reconstruction as the feature extractor to obtain RA-related features for DL interpretation. Trained by comparing with the true label and the so-called 'label consistency' between the original and augmented image, the classification or prediction result for a specific task is produced as output.

The next four subsections will successively introduce the preprocessing, the

backbone model, the self-supervised pretraining, based on the anatomical consistency, and the contrastive loss function, based on label consistency of the same samples.

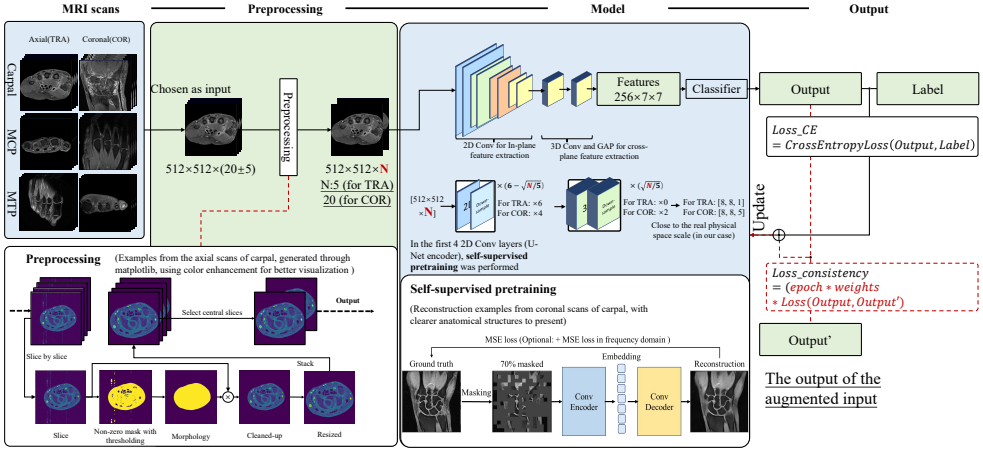


Figure 3.2: Overall workflow. The training framework includes three main steps: (a) Preprocessing; (b) Model and its self-supervised pretraining; and (c) a self-contrastive loss function. COR: coronal; TRA: transversal (axial); MSE: mean squared error; Conv: convolutional layer. The value N, which represents the number of slices, is set to be five for TRA scans and twenty for COR scans to improve the information density based on our previous study [36][85].

3.3.2 Preprocessing

The variety and complexity of anatomical and pathological structures and spatial placement of hands and feet amplify the overfitting of deep learning models in these RA-related applications. To overcome these challenges, images were first preprocessed by background removal, resizing and intensity normalization. In addition, central slices were selected from the 3D axial (TRA) scans as the input to improve the efficiency of model training and increase the density of RA-related information, as the proportion of background (air) to foreground (hand or foot) on the top and bottom of each scan is generally high.

Images were first thresholded at 10% of the maximum intensity value in the image and processed through morphological opening and closing operations [73] to obtain the masks of targeted foreground objects [74]. The threshold was fixed according to a small subgroup of this dataset, and visually checked on the training set. Some thresholding algorithms such as OTSU might improve the thresholding process and help to generalize the preprocessing to other datasets as some studies stated the OTSU outperforms fixed thresholding [86]. However, in this study, fixed thresholding outperformed OTSU in distinguishing the foreground and background in most cases. The automatic thresholding methods could cause information loss due to over-thresholding in the targeted anatomical structures (some examples were

presented in the supplementary materials). In the cases, where the fixed thresholding fails, the main challenge is that some irregular gradient intensity changes blur the anatomical borders of the wrists, MCPs and especially the MTPs. In this situation, the definition of the anatomical borders that have intensities equal to the backgrounds becomes the primary problem. Some more advanced or customized thresholding algorithms could solve this problem for more general use, which requires further investigation.

After the thresholding to exclude the backgrounds, the images were resized to similar sizes without changing the aspect ratio, to minimize the variety of foreground object sizes. Subsequently, these images were normalized individually and slice-by-slice to zero mean and unit variance, with a 95% clipping to avoid over-normalization caused by the extremely high values from inflamed areas.

Moreover, the size and foreground-to-background ratio of 3D MRI scans reduced the efficiency and amplified the difficulty of model training. Therefore, based on previous work [36], and the observation that the foreground-to-background ratio decreases significantly with increasing distance from the central slice, the central five slices were selected as input instead of the whole 3D scans. Here, the central five slices were defined as the slices with the largest sum of non-zero mask area in the previous masking process. The number of central slices (N in Fig. 3.2) was determined based on pre-experiments from our previous study [85], in which five central slices could perform as well as using all 3D scans in TRA. For the coronal (COR) scans, the variety of spatial placements and the irregular gradient intensity change in the scans make it difficult to select the central slices automatically. Therefore, in this study, the N for TRA scans is set to five and for COR is set to twenty.

3.3.3 Backbone model

Because the COR and TRA scans describe the same anatomical sites (carpal, MCP, and/or MTP joints), but with different resolutions in each direction, the model architecture was designed to adapt different sizes of images from TRA and COR scans and then output the extracted features in a fixed shape. Considering the limited number of samples and different task complexities, we implemented a model transferred from the basic U-Net [78] encoder as the backbone, for potential pretraining and transfer learning. The model architecture contains two main parts: (1) an encoder that contains both 2D and 3D convolutional (Conv) layers to output features of a same size for different scans; and (2) a standard dense layer as a classifier. The encoder is formed by sequentially stacking 2D (kernel size:[1, 3, 3]) and 3D (kernel size:[3, 3, 3]) Conv layers, with the same hyper-parameters as the basic U-Net. Inspired by the resampling process in nnU-Net [74] that standardizes the input size at the image level, the number of 2D and 3D Conv layers is set to accommodate samples with

different input sizes and output a fixed number of features to perform a feature-level standardization. The reasons for using the U-Net encoder, with a few changes, as the backbone of 2D Conv parts are: (1) Most advanced model architectures and functional modules require a large amount of training data, which is not available; (2) U-Net encoders are simple to be implemented and reproduced for both researchers and users; (3) U-Net encoders, which have been widely used in both natural (known as the VGG encoder) and medical imaging field (encoder part of a U-Net), is naturally convenient for pretraining through transfer learning or self-supervised training (see next section).

The encoder produces features as input for a subsequent classifier by simply stacking an adaptive pooling layer and three dense layers. The configuration of the whole model architecture and training can be found in the supplementary materials.

3.3.4 Anatomical consistency and self-supervised pretraining

For the objects (wrist, MCPs or MTPs), MRI scans from different subjects (patients) share similar anatomical structures and their spatial placement (e.g., carpal bones, ulna, radius) with only a few variations caused by individual anatomical and pathological differences. These similar structures and spatial placements are called "anatomically consistent" information in the whole dataset, which is common knowledge for clinicians, yet not fully utilized in a DL model as labels usually do not contain this prior information.

To pre-train the model when samples are limited, a self-supervised method called masked auto-encoder (MAE) [83] was employed. Compared to natural images analyzed by self-supervised methods, our number of samples is limited, yet the structures of human hands and feet are anatomically consistent, which could be learned by models and used as prior knowledge. For example, the number of bones or the existence of inflammation around tendons in the wrist could be a hint to finding bone marrow edema and tenosynovitis, respectively, that are related to RA. In a previous study, a self-supervised pretraining strategy was explored in medical imaging by [87] with a series of augmentation methods, proving the potential of applying self-supervised reconstruction as pretraining to 'warm up' the model with a similar task. Compared to the method of Zhou, which requires models to reconstruct original images through differently-augmented images, the training strategy of MAE is simpler, and more efficient by reconstructing original images from randomly-masked images. Since valid results with good generalization ability have been achieved by MAE in natural imaging fields, we extended it from a process based on Transformers [84] to a process based on the U-Net that contains less parameters, which is more suitable for medical imaging, but with the same principle of reconstruction from masked images to learn underlying semantics.

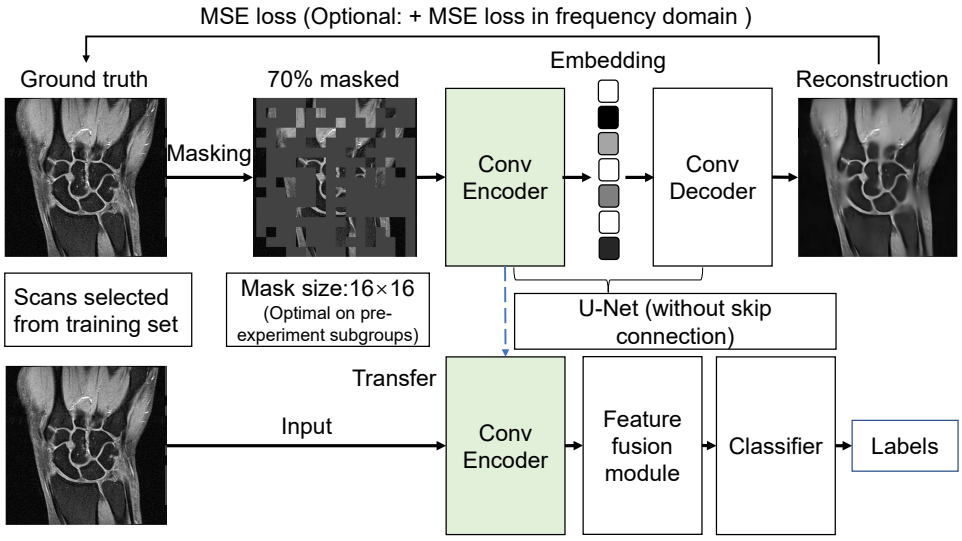


Figure 3.3: The workflow of the self-supervised reconstruction. The mean squared error (MSE) loss contains two parts, the MSE loss of spatial domain and frequency domain.

Fig. 3.3 presents the basic idea of building a self-supervised pre-training process on the U-Net encoder. To define the reconstruction task, 70% of the input image was masked by patches (16×16 pixels), which were distributed randomly. Using skip connections in this reconstruction-based pre-training would introduce a risk of having the model not learn the underlying patterns or anatomical structures at a high level, but copy and paste over the epochs to get the reconstruction in the corresponding areas. Therefore, we expected that the encoder would learn more underlying patterns of anatomical structures of human hands and feet, if skip connections are removed, to avoid information leakage through these shortcuts. Although models with skip connections could performed quite well on the training set, pre-experiments [85] showed that they failed to achieve good reconstructions in validation sets that are not involved in the training process.

Compared to Transformers-based MAE, CNN-based MAE encounters more quality problems because patches contain both masked and unmasked pixels, which makes it difficult to reconstruct high-quality images. Therefore, in addition to the pixel-to-pixel MSE loss of the reconstruction and original images and the MSE loss based on the frequency domain was combined to improve the reconstruction results. The loss is given by: $loss = alpha \times MSE(output, GT) + (1 - alpha) \times MSE(freq(output), freq(GT))$, where $output$ refers to the prediction of the models, and GT refers to the ground truths. The hyper-parameter $alpha$ was set to 0.8 for maximum convergence speed in this work,

and more details for the relationship between the loss function, epochs required for convergence and the $alpha$ can be found in the supplementary materials.

3.3.5 Label-consistency loss function

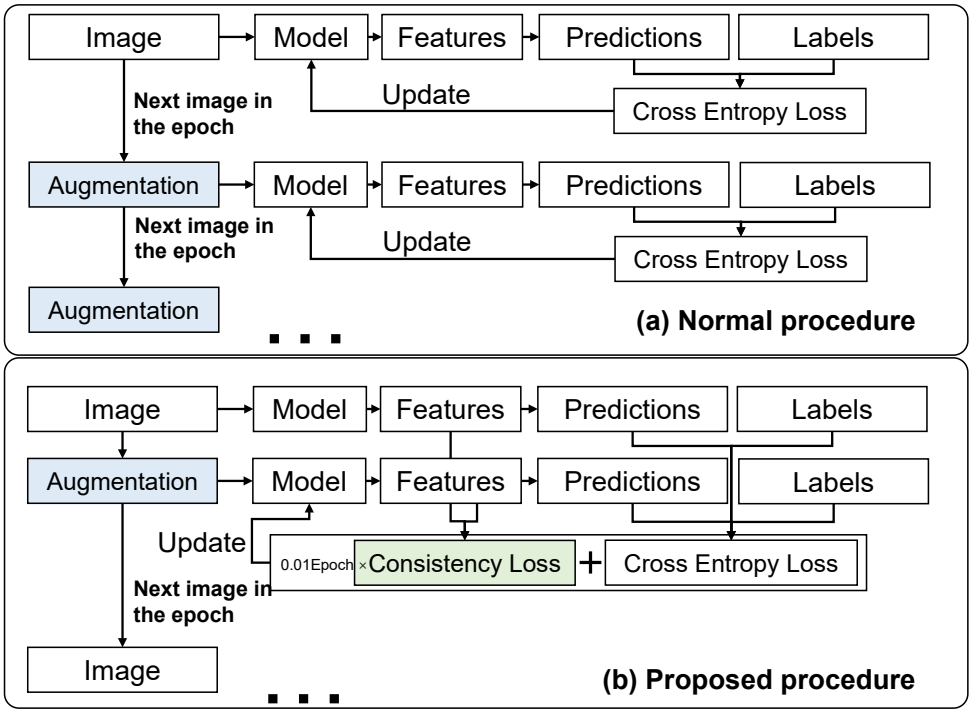


Figure 3.4: a) the normal procedure of using augmentation and b) the procedure when the consistency loss function is added to the training process.

Similar to other DL methods in medical imaging, basic data augmentation was applied to overcome overfitting. Besides classical ways of augmentation, we took a different approach to use data augmentation, inspired by contrastive learning [88, 89, 90], to maximize the value of the shared information in the original image and the augmented one. This could help model training to focus on the unchanged RA-related information, by excluding the impact of spatial placement and added noise more efficiently.

We propose a loss function, called the 'label-consistency' loss function, to take advantage of the fact that the properly-augmented and original input data share the same label and disease severity, which is the so-called "label consistency" of RA information. The assumption for this loss function is that the RA-related information remains unchanged during the augmentation, as a defined spatial transformation and added Gaussian noise will not remove any lesions or anatomical structures. To comply with this assumption, the augmentation is limited to avoid large transitions

or extensive cropping, and a margin (0.05) for the loss between the output is set to leave some space for accidental cut-offs by augmentation. The consistency loss function is aiming at minimizing the differences in the output logits between the input and augmented image, in one training epoch, while the cross-entropy loss function is trying to maximize the differences between different classes. To stabilize the training process, an epoch-dependent weight is added to minimize the impact of this extra loss function at the early stage of training and increase the impact at a later stage. The consistency loss function is given by: $loss = CrossEntropy(output, GT) + i \times w \times Margin(MSE(output, output_a))$. The cross-entropy part of the equation is the same as in standard classification tasks, where $output$ represents the output logit of the model, and GT refers to the ground truths. The second part represents the consistency loss, where i refers to the index of the training epoch in the range from zero to the maximum number of epochs-1; and w is a hyper-parameter used to control the weight of consistency loss. At the last epoch, the product of i and w will reach 1.0 to gain an equal effect as the cross-entropy loss. This gradually rising weight is to avoid affecting the direction of model learning in the early stages of training and enabling the model to converge. In our experiments, the range of w was set from 0.005 to 0.01 as the total number of epochs varies from 100 to 200 because of task differences; $output$ and $output_a$ refer to the output logits of the model with as input the original image and its augmented version, respectively. The $Margin()$ function leaves space for small values of MSE loss that might be caused by accidental occlusion of information by augmentation. Fig. 3.4 presents the workflow after adding the loss function.

3.3.6 Class activation mapping

The CAM technique [91, 58, 55] is one of the most common techniques to open the deep learning black box. Since in our case, classification mainly applies to the center of the images, we applied the pixel-to-pixel calculation of the original gradients instead of the average gradients in Grad CAM. Moreover, to fully reflect the model's judgment criteria, we retained the negative parts, which are usually removed in standard CAM methods, in which only the positive regions are presumed to represent objects appearing in the background. As the regions with negative values in saliency maps can represent normal objects that decrease the confidence of reporting early RA, we preserved them, resulting in activation values in the background (air) greater than zero, but still representing "no contribution" . Consequently, regions with activation values lower than that of air represent a negative contribution to the targeted label. The results of CAMs can be found in the next section, which illustrates the focus of the models for RA classification/prediction.

3.4 Results

3.4.1 Evaluation principles

The area under the receiver operating characteristic curve (AUROC) was employed as an evaluation metric for 5-fold cross-validation and during testing, calculated from the datasets with the labels of EAC, CSA and healthy controls for the first two tasks and the labels of RA and non-RA in the third and fourth tasks. The standard deviation (SD) of a given AUROC was calculated during 5-fold stratified cross-validation, as presented in the following tables. Moreover, to avoid AUROC being overly optimistic due to data imbalance, the number of samples for each class in the test- and validation-set was kept similar. All experiments were executed on an RTX6000 GPU from Nvidia, with PyTorch 1.12 https://pytorch.org/ on Python 3.9 https://www.python.org/ and SciPy 1.7 https://scipy.org/. All experiments were executed ten times with random seeds, and the results of the models with median performance were presented. Details on the configurations used for self-supervised pre-training and fine-tuning can be found in the configurations-section in the supplement.

3.4.2 Reconstruction examples from self-supervised pretraining

Fig. 3.5 provides some examples from the self-supervised pre-training process. The first four convolutional layers, with a similar structure as the U-Net encoder, were trained to extract features from 70% masked images and reconstruct the original images with a decoder.

As shown in Fig. 3.5(a), the model can predict unseen anatomical structures in the masked regions (highlighted by the red boxes) when 70% of the image in the test set was masked, without any labels or prior knowledge. When the 99%-masked images were fed into the models trained on 70%-masked images, as shown in Fig. 3.5(b), the model still grasped the basic concepts and structures of carpals although it was unable to predict the whole anatomy due to insufficiency of information. These results prove that the pre-trained models have learned some basic anatomical knowledge of carpals without any label.

3.4.3 Overall performance on the four tasks

The overall performance of the proposed method on all four tasks can be found in Fig. 3.6. The details of the input and results can be found in Table 3.2. The first two classification tasks served as pretraining for the RA classification/prediction tasks. In this phase, the proposed models were trained on the carpals, MCP and MTP, separately, and a combination of these scans. Due to the overfitting problem of MTP-based models, the combination of three anatomical sites failed to reach competitive results, therefore

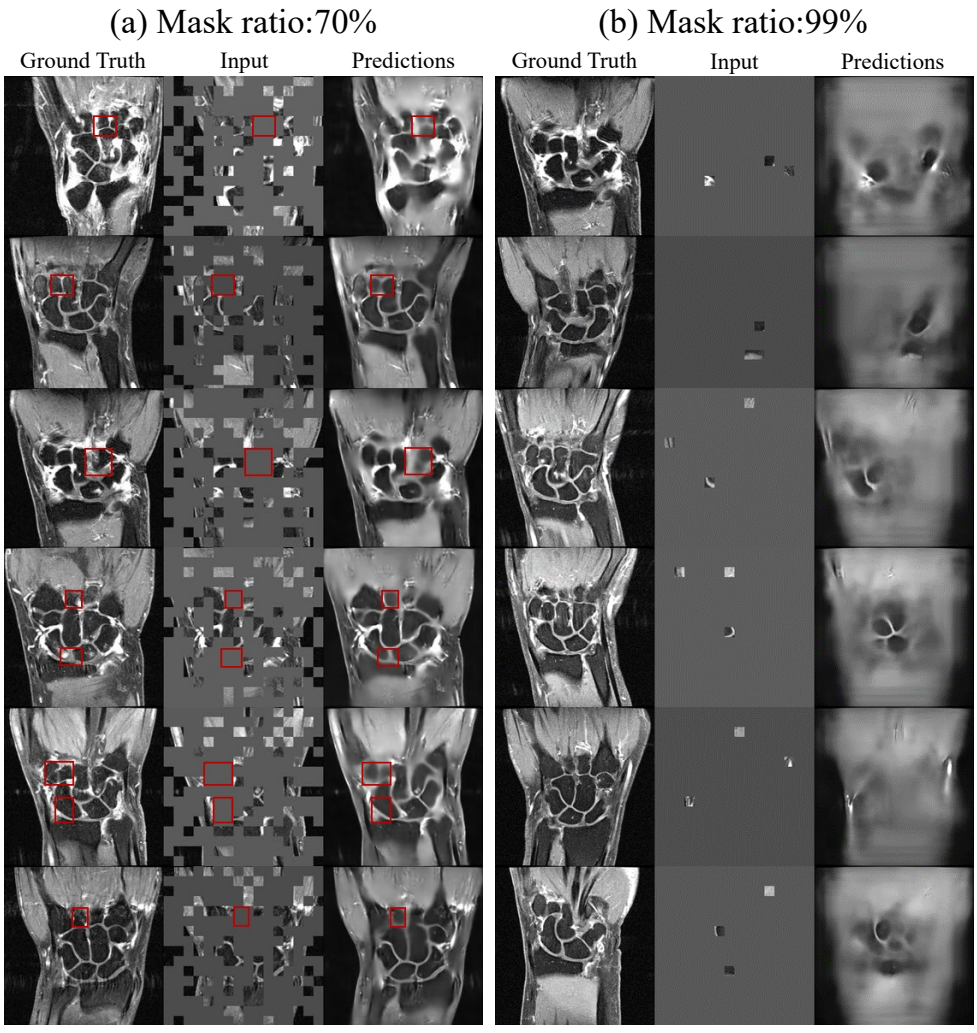


Figure 3.5: Reconstruction examples from the model trained on 70% masked input, (a) test on 70% masked scans from the held-out test set. (b) test on 99% masked scans from the held-out test set.

these are not presented in Fig. 3.6 and Table 3.2. Consequently, the models for MTP-based RA prediction were pre-trained by the reconstruction models only.

For classification tasks that distinguish early-onset arthritis or clinically suspect arthralgia from healthy controls, the models achieved AUROCs of over 0.8 on cross-validation and close level on the held-out test set. However, the performance dropped from around 0.65 to 0.7 for the third classification task. This mainly originates from the difficulty of distinguishing RA from other arthritides, as inflammatory areas may

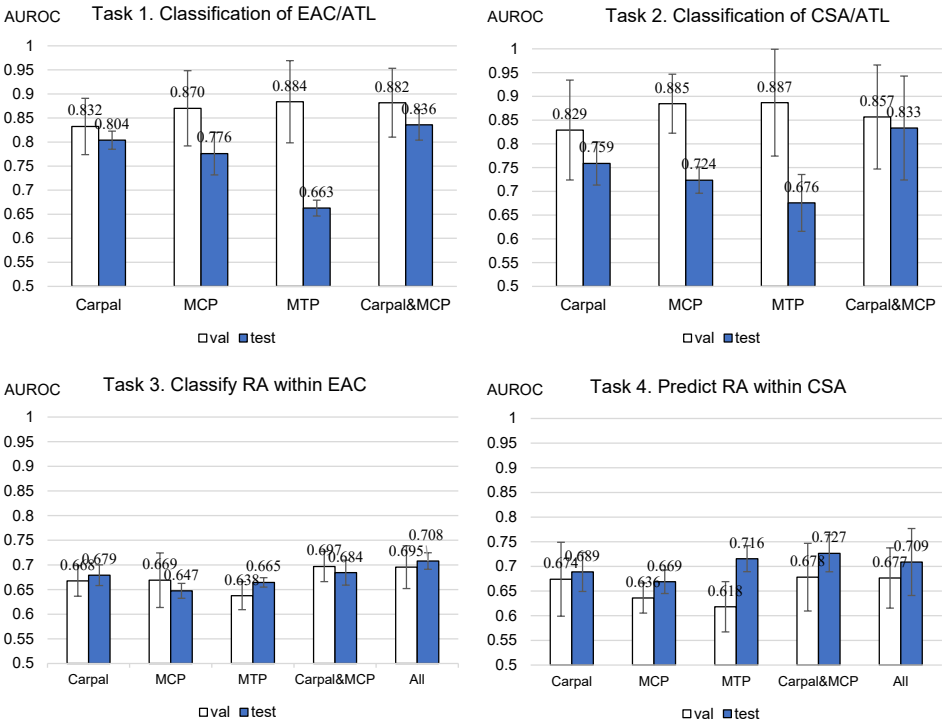


Figure 3.6: Overall performance on the four tasks. The solid bars represent the results of the test set, while the hollow bars represent the results of the cross-validation. In most tasks, the ensemble models using the combination of carpals and MCPs obtained the highest AUROCs. The MTP-based models suffered overfitting and performed poorly on the EAC and CSA classification tasks, because of confounders related to the stability of the MRI scanner over time, which were analyzed in the supplementary materials.

significantly contribute to distinguishing arthritides, yet are common in both RA and other arthritides. For the prediction task, which is more challenging, the performance of the DL model is comparable to statistical analysis on clinical variables [10] (AUROC equal to 0.74, with different validation set.).

3.4.4 General improvements compared to baseline models

Because of the complexity of tasks and the many combinations of inputs, the comparison results were given based on the best results available on each task without considering the remaining combinations of inputs. Due to the lack of related studies in this field, we re-implemented the ResNet18/34/50/101/152 and VGG11/13/16/19

Table 3.2: Overall performance on each task with different inputs

Task	Input	AUC (± Std.)	
		Validation	Test
Task 1: EAC vs ATL	Carpal	0.832 (±0.058)	0.804 (±0.019)
	MCP	0.870 (±0.078)	0.776 (±0.044)
	MTP	0.884 (±0.085)	0.663 (±0.016)
	Carpal + MCP	0.881 (±0.072)	0.836 (±0.032)
Task 2: CSA vs ATL	Carpal	0.829 (±0.105)	0.759 (±0.045)
	MCP	0.885 (±0.062)	0.724 (±0.028)
	MTP	0.887 (±0.112)	0.676 (±0.059)
	Carpal + MCP	0.857 (±0.110)	0.833 (±0.109)
Task 3: Classify RA within EAC	Carpal	0.668 (±0.031)	0.679 (±0.021)
	MCP	0.669 (±0.055)	0.647 (±0.015)
	MTP	0.637 (±0.028)	0.664 (±0.009)
	Carpal + MCP	0.697 (±0.031)	0.684 (±0.025)
	Carpal + MCP + MTP	0.695 (±0.043)	0.708 (±0.017)
Task 4: Predict RA within CSA	Carpal	0.674 (±0.075)	0.689 (±0.039)
	MCP	0.636 (±0.031)	0.669 (±0.024)
	MTP	0.618 (±0.051)	0.715 (±0.026)
	Carpal + MCP	0.678 (±0.068)	0.726 (±0.037)
	Carpal + MCP + MTP	0.676 (±0.061)	0.708 (±0.068)

Table 3.3: Best results available based on different models on each task

Task	Models	Input	AUC (± Std.)
Task 1: EAC vs ATL	ResNet34	Carpal + MCP	0.721 (±0.059)
	VGG16	Carpal	0.732 (±0.026)
	ViT	Carpal + MCP	0.660 (±0.092)
	consistVGG	Carpal + MCP	0.836 (±0.032)
Task 2: CSA vs ATL	ResNet34	Carpal + MCP	0.637 (±0.060)
	VGG16	Carpal	0.673 (±0.026)
	ViT	Carpal + MCP	0.612 (±0.060)
	consistVGG	Carpal + MCP	0.833 (±0.109)
Task 3: Classify RA within EAC	ResNet34	MTP	0.547 (±0.061)
	VGG16	Carpal + MCP	0.631 (±0.089)
	ViT	Carpal + MCP	0.612 (±0.060)
	consistVGG	Carpal + MCP	0.708 (±0.017)
Task 4: Predict RA within CSA	ResNet34	MCP	0.526 (±0.047)
	VGG16	MCP	0.614 (±0.108)
	ViT	Carpal + MCP	0.579 (±0.041)
	consistVGG	Carpal + MCP	0.726 (±0.037)

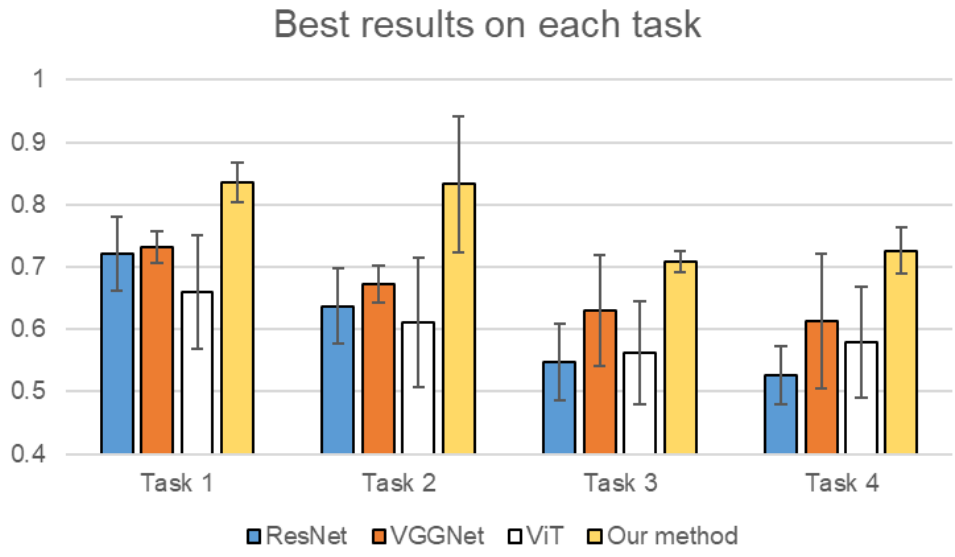


Figure 3.7: Best results from each method applied on the test set. The blue bars present the results of ResNet34 [92] on test set in each task, while the orange bars are the results of VGG16 [93]. The best result of Transformer-like models is presented in white bars, using a ViT [94]. These models achieved the best performance of their kinds (ResNets/VGGs) in the four tasks. The results of the VGG models trained with the consistency-based methods (consistVGG16) are given by yellow bars.

with or without the attention module and dense block as the baseline for these four RA classification/prediction tasks. These models have been the most widely used backbone architectures for DL-based medical tasks. For example, in breast MRI [95], Alzheimer's MRI classification [96] and disc degenerative disease based on MRI [97], when LSTM and some other structures or modules (e.g. attention modules [98]) were introduced because of specific data characteristics, the CNN backbones remained to be ResNets and VGGs. Meanwhile, the comparison with these baseline models could more clearly prove the effectiveness of the proposed strategies.

ResNet3D in [99], which is the closest study to our task, was also implemented. However, the input resolution and tasks were different, making it fail to perform these tasks and cannot outperform the backbone (ResNet). For more advanced models, like Transformer, we were not able to train models because they are too data-hungry. Apart from the baseline models, we also applied the widely-used lightweight models such as MobileNet [100] and MobileViT [101] to investigate different types of models, the results can be found in the supplementary materials. Similar pre-experiments were also implemented to validate the effectiveness of other modules such as attention

modules, multi-scale processing and multi-task training. However, all these attempts presented no statistical significance in improving the model performance.

As shown in the Fig. 3.7, compared to all the baseline CNN models and a ViT-B [94] model, our models present substantial increases of AUROCs in the RA classification/prediction tasks. Especially in the CSA-related tasks, our models achieve significant AUROC improvements over 10% are achieved. Meanwhile, the MRI scans of carpals and MCPs appeared to be the most informative for the RA-related tasks. More details of the best models can be found in Table 3.3 and the performance of lightweight models can be found in the supplementary materials.

3.4.5 Ablation study of each proposed component

Fig. 3.8 presents the results of ablation experiments applied on the classification Task 1 and 2, which contain all the proposed strategies in the training process, based on the MRI of carpals. The self-supervised pretraining and consistency loss function contributed most to the performance, especially for the CSA classification task, while the model architecture and pre-processing also have a clear impact on the AUROC. The contribution of each strategy varies because of the variation in input materials, yet delivered a clear message that the performance of deep learning models in medical fields is highly dependent on the training strategies.

3.4.6 Saliency maps generated by CAM

In Fig. 3.9, examples were randomly selected and organized based on the output confidence of NNs, as can be seen in the axes. As the segmentation ground truth of lesions on our dataset is not available, most tasks based on datasets with pixel-level ground truth will be turned into segmentation tasks instead of classification or prediction. It can be found in all the figures that with the increase in confidence, the number of high-intensity pixels increases. We also applied an algorithm to merge the saliency maps generated from different nodes of the neural networks and normalized them to the range of 0 and 1. Therefore, saliency maps were normalized through the max-min normalization, while the scores of the air always represent the correlation of zero. This leads to the high intensity of air caused by the normalization in some images because the scans contributed very little. That's the reason for the high values in the air in saliency maps from Task 4.

3.5 Discussion

On average, our models achieved AUROCs of 83% and 70% in the RA classification/prediction tasks with the proposed strategies.

From a clinical perspective, the DL models obtained reasonable results in distinguishing EAC, CSA and healthy controls, indicating the potential of applying DL models

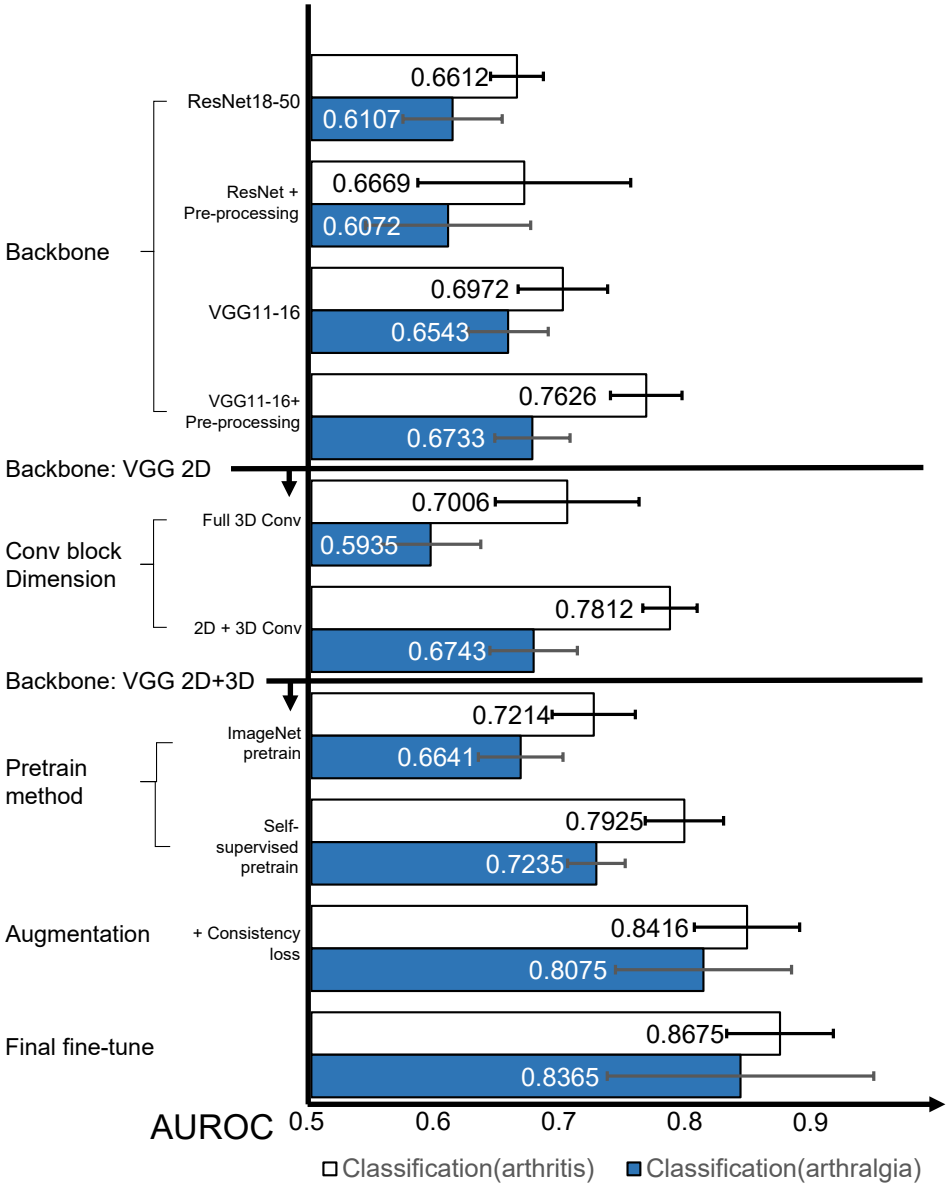


Figure 3.8: Contribution of each proposed strategy.

to assist in detecting the development of arthritis and CSA. However, the performance of DL models in distinguishing RA from other arthritides requires further investigation and improvement to assist in arthritis identification. The difference between the first two tasks and classification of RA demonstrates that the models rely on inflammatory signs. The performance of the models in RA prediction, with AUROCs of 70% on this

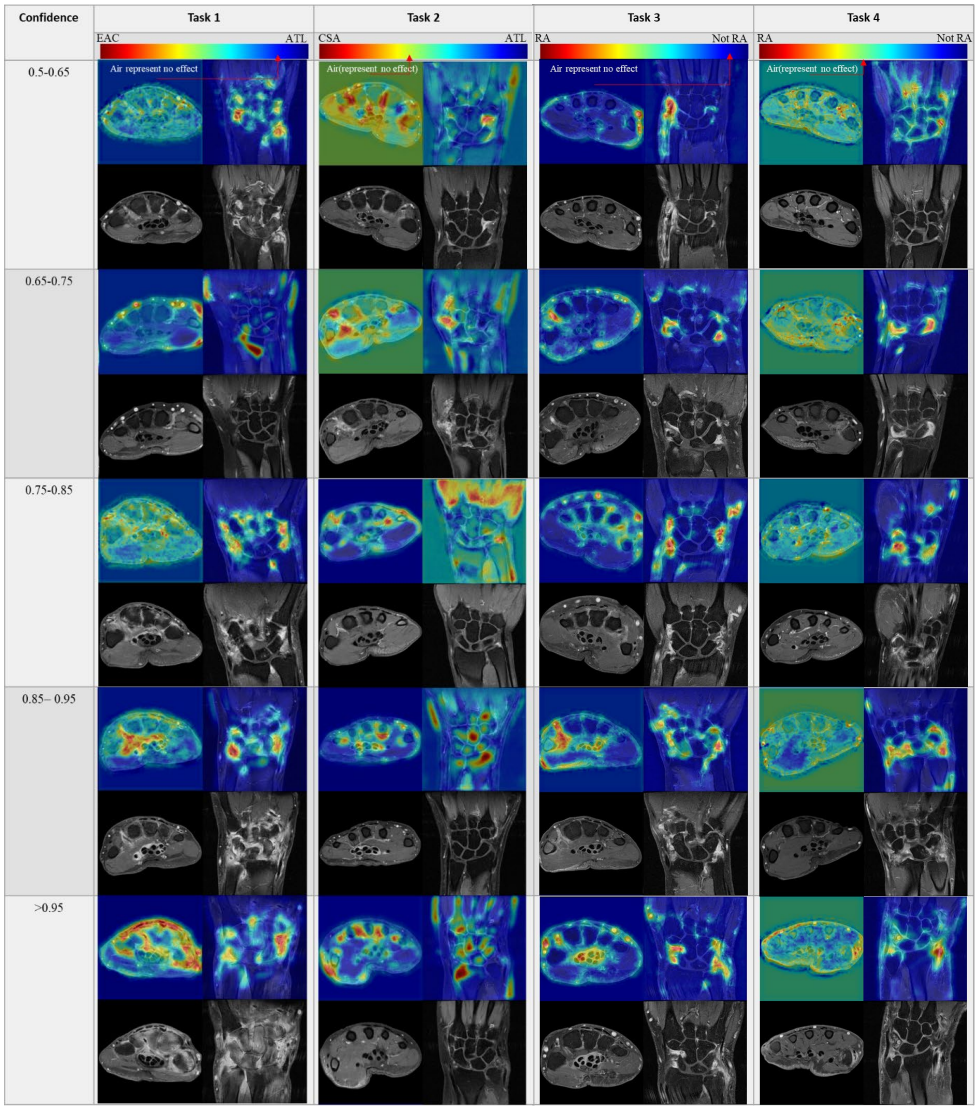


Figure 3.9: Saliency maps for the four tasks. The closer the color is to red, the more the pixel contributes to the model decision for EAC, CSA or RA. Compared to the original Grad CAM, the background in the saliency maps generated in our method has a value larger than zero (yet still represents background contribution), as the threshold of no contribution to the targeted category, pixels with values less than this threshold have negative contributions, while the opposite ones have positive contributions to the model decision for this category.

challenging task, is comparable to the performance by statistical analysis of clinical variables, demonstrating the potential of using DL models to search for early RA signs.

According to the saliency maps, inflammatory signs appear to be the most contributing factor in current models, which is consistent with clinical knowledge. However, due to the limited dataset capacity, data imbalance and lack of other RA datasets for general validation, clinical application requires further investigation and validation.

From the perspective of method, the consistency-based strategies significantly improved the overall performance of the baseline model on all tasks (See Fig. 3.7). The self-supervised reconstruction, based on the consistency of specific anatomical structures, provides a pre-training method for CNNs when the amount of data is limited and when there is no similar dataset available for pre-training. As far as we know, our method is the first DL-based method for the detection of early signs of RA from MRI. Most previous studies related to RA were based on Ultrasound [102, 45] or X-rays [103] and were focusing on predicting visual scores. Moreover, most other studies for MRI-based diagnosis are based on the combination of prior knowledge and the variants of standard ResNets and VGGs. These backbone models have therefore been included in our method comparison, as specific prior knowledge in these fields cannot be transferred. Compared to other methods, our methods focus more on learning from the information consistency, namely unchanged anatomical structures, and labels during processing (augmentation and DL-based feature extraction). These strategies are transferable, as they are generally not dependent on any specific prior knowledge. However, they are very dependent on the reconstruction quality and augmentation methods and therefore rely on anatomical consistency and disease severity consistency during the spatial transformation.

Although there is still a gap with the most experienced clinical expert, this is already a big step toward to fully-automated prediction of early RA, and the most advanced approach for automatic detection of early RA so far. Meanwhile, this work also indicates the existence of early signs of RA in MRIs without prior knowledge from rheumatologists, which could serve future studies that use this modality. Moreover, the performance of deep learning models can be further improved as several impactful factors can be studied. First, for preprocessing, the augmentation methods were limited to small-scale spatial transformation and noise addition that would not change the labels. To help models overcome some artifacts (e.g. caused by fat suppression errors), the models can be improved if we can mimic these artifacts and use them in augmentation for training, approaching the expert level of robustness against image degradation. Moreover, we selected central slices automatically based on non-zero masks, which can be sometimes mistaken. Anatomical knowledge and advanced segmentation may play a role here to improve it.

Meanwhile, apart from labels, deep learning models are independent of other expert knowledge, we expect more information than the prediction of labels. With a verified visualization method that can test the reliability of models, the deep

learning models can also serve as a way of exploring inflammatory signs of RA that have not been considered by clinical observers but could still be relevant to RA development (i.e., hypothesis-free interpretation). Combined with the current visualization methods, as shown in Fig. 3.9, the saliency maps can already illustrate some potential regions where some early signs of RA exist. With our ongoing studies on improving visualization methods, it has become feasible to generate saliency maps and find early signs of RA. This may give a different perspective for studying RA or finding potential image biomarkers for early RA.

3.6 Conclusions

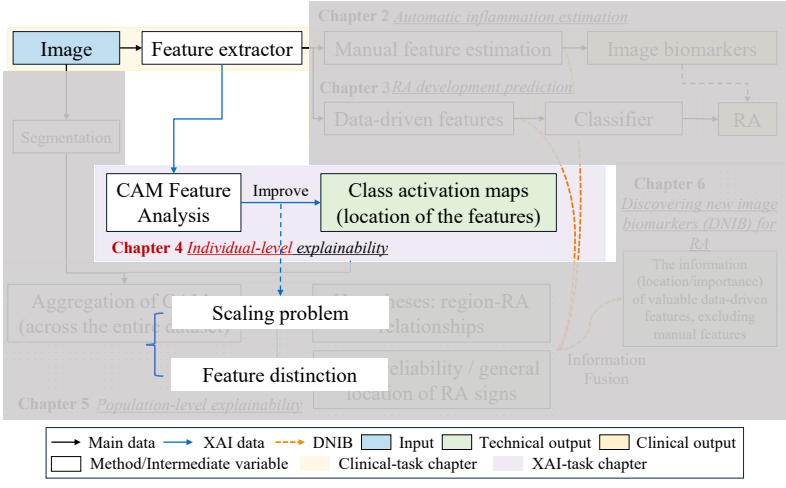
The proposed method in this chapter is the first DL-based method for detecting the early signs of RA from MRI. The models, based on the proposed consistency-based strategies, succeeded in all four RA classification/prediction tasks. This indicates the existence of early signs of RA in MRIs and it demonstrates the potential of DL models in RA-related research. The proposed model could serve as an initial DL benchmark in RA prediction based on MRI and indicates the ability of DL to assist RA analysis and finding early signs of RA in MRI scans with the visualization method, contributing to both technical and clinical RA studies in the future. The results of this chapter also set up the fundamental basis of the following chapters.

Acknowledgment

This work is supported by the Netherlands Organization for Scientific Research (NWO, TTW 13329), the European Union's Horizon 2020 research and innovation programme (No.714312) and the China Scholarship Council (No.202108510012).

4

Feature analysis for proper intensity scaling and feature distinction in class activation maps



This chapter was adapted from:

Yanli Li, Denis P. Shamonin, Tahereh Hassanzadeh, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "Feature analysis for proper intensity scaling and feature distinction in class activation maps." (*submitted*)

Abstract

Understanding the decisions of deep learning (DL) models is crucial for their acceptance in risk-sensitive applications. Class activation maps (CAMs) are commonly used in image analysis to visualize model reasoning by generating attention maps where signal intensities represent contributions to outputs. However, existing CAM algorithms focus on optimal weight design and salient feature layer selection, neglecting two key limitations in population-level interpretation: (1) lack of an appropriate intensity scale for proper interpretation and quantitative analysis, and (2) inability to identify which features within selected layers predominantly drive model reasoning. These gaps can lead to miscorrelations between CAMs and model outputs, causing erroneous interpretations, while restricting CAMs to case-specific visual inspections. We propose a framework to statistically analyze DL-extracted features at a population level, determinizing feature contributions for global intensity scaling and within-layer feature distinction. The global intensity scale standardizes CAMs, achieving high R with model outputs. Within-layer feature distinction identifies overfitting, confounding factors, outliers, redundancies, and principal features. Applied to eight datasets, including five medical imaging datasets, this framework improved Rs between CAMs and outputs by 10.7 – 64.2%, achieving near 100% consistency. Furthermore, distinguishing principal features (5 – 25% in the selected layer) produced CAMs of equal quality while maintaining model output accuracy. This method, relying on minimal assumptions, enhances CAM-model consistency and broadens CAM applications by enabling standardized interpretation and deeper feature-level insights.

4.1 Introduction

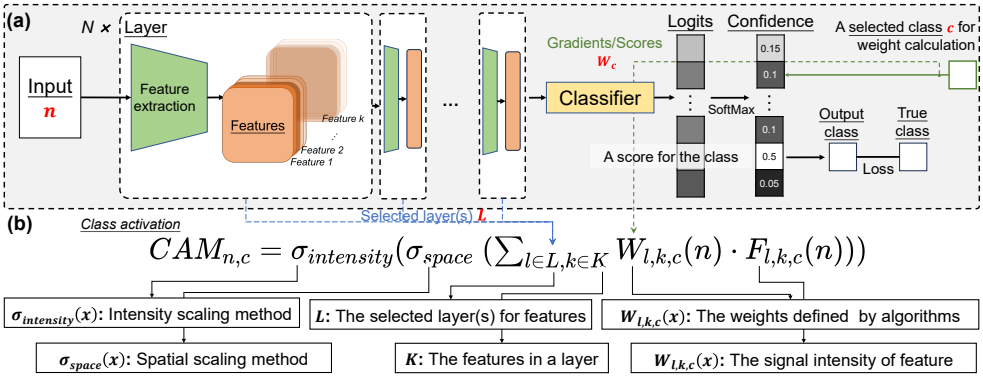


Figure 4.1: Summary of the concepts, calculation process and terms involved in generating CAMs. (a) The flowchart at the top shows the general DL workflow, including: (1) *input* is fed into multiple consecutive layers of feature extraction modules, obtaining (*features*) given by these layers; (2) the features in the last layer(s) of a model are fed into the classifier to generate *logits*, which represent the scores of different output classes; (3) *confidence* is then obtained by applying softmax function to *logits*; (4) based on the *confidence* to get the *output class* and compare with the ground truth *true class*. (b) The formula in the middle summarizes the calculation and elements used in current CAM algorithms to generate a $CAM_{n,c}$ for a selected class c (often default to be output class) and an input n (from the set of all inputs N_c): the designed weights $W_{l,k,c}$, a selected layer l from the set of all selected layer(s) L, spatial scaling $\sigma_{space}(\cdot)$, individual intensity scaling $\sigma_{intensity}(\cdot)$ and selecting the features k (from the set of all features K within the selected layer(s). Underlining highlights the important terms that are consistently used in this study and have specific meaning, including: (1) *input* – input of DL models; (2) *layers* – the functional layers of DL models; (3) *features* – the values that go through a specific path (feature extraction process) during the feedforward process in a DL model, also known as latent or embedding; (4) *logits* – the raw outputs of a DL model before any transformation; (5) *confidence* – *logit* after a Softmax function; (6) *output class* – the prediction/classification result according to the *confidence*; (7) *true class* – the ground truth; (8) the *selected class* – a class selected for generating the CAM for that specific class; and (9) class *activation* – signal intensity in CAMs.

Validating the reliability of deep learning (DL) models and understanding how these models analyze images are essential requirements before DL models can be deployed in risk-sensitive areas [104, 105, 106, 107]. In addition to validation based on large-scale data, it needs to be confirmed that DL models are actually reasoning based on trustworthy evidence and convincing features [50, 51, 52, 53, 54, 55, 56]. In image analysis, this confirmation specifically refers to ensuring that the focus of a model on the input images should be spatially and semantically consistent with the focus of

expert knowledge.

To ensure this consistency, class activation maps (CAMs) and their variants are the most representative methods to reveal the focus of DL models and compare this focus with expert knowledge, since they generate attention maps, in which the signal intensities represent the importance (class specificity or contribution) of a region to the output of the DL model. Especially in image analysis, CAMs became one of the most popular methods due to its high computational efficiency and intuitive ways of displaying, compared to other interpretability methods, such as LIME [108, 109, 110, 111], global interpretation [112, 113, 114] and other interpretation methods [115, 116, 117, 118, 119, 120].

The family of CAMs shares a similar calculation that is dependent on two main components – features and weights. Features refer to the outputs of certain layer(s) in DL models, which contain spatial information for displaying and activation intensities during reasoning. The weights represent the importance of these features to the output of DL models, similar to the weights for variables in statistical regression models. In Fig. 4.1, we summarized the calculation of the most prevailing CAM algorithms and identify certain terms used in the calculation process, including the original CAM [55], Grad CAM [58], Grad CAM++ [59], Layer CAM [121], xGrad CAM [122], Score CAM [123], Smooth Grad CAM++ [124], SS-CAM [125], IS-CAM [126], Eigen CAM [127], Multi-CAM [128] and Fusion CAM [129].

These algorithms provide an important insight into the decision making process, achieving promising performance in the fields like weakly supervised object detection [131, 132]. However, two important issues have not yet been considered when generating CAMs – there is no appropriate intensity scale and they do not give insight into which features within the selected layer(s) substantially contributed to the model' s reasoning. As shown in the formula in the middle of Fig. 4.1, five elements are indispensable for CAM calculations: (1) defining weights, (2) selecting layer(s) that contain features, (3) scaling spatially, (4) intensity scaling and (5) selecting features within the selected layer(s). While existing algorithms are designed to optimize the definition of weights and the selection of the layer(s), few investigations were conducted on the remaining three elements as shown in Tab. 4.1. It is generally reasonable to not pay much attention to optimizing spatial scaling [133] - the architecture of mainstream DL models, such as translation equivariance in convolutional neural networks (CNNs) and the Transformer encoder's feature of preserving input dimensionality throughout its layers, help to simplify spatial scaling. However, the remaining two elements, intensity scaling and distinguishing the most important features within the selected layer(s), could have a considerable impact on the accuracy and applicability of CAM algorithms.

Table 4.1: The weight definition and feature/layer definition in some CAM algorithms. While the weight definition varies in each algorithm, the features in the selected layer(s) are typically all adopted without distinction and the scaling are commonly individual intensity scaling.

Method (Gradient)	**Weight definition**	**Feature and layer definition**	**Scaling method**
Grad-CAM[58]	Average of gradients of each feature map: $W_{l,k,c} = \frac{1}{\lvert i \times j \rvert} \sum_i \sum_j \text{gradient}_{i,j,l,k,c}$	All features from the last feature extraction block	Individual intensity scaling
Grad-CAM++[59]	Pixel-weighted average of gradients: $W_{l,k,c} = \sum_i \sum_j \alpha_{i,j}^{k,c} \cdot \text{ReLU}(\text{gradient}_{i,j,l,k,c})$	All features from the last feature extraction block	Individual intensity scaling
Layer-CAM[121]	Pixel-level gradients from each stage: $W_{l,k,c} = \text{gradient}_{i,j,l,k,c}$	All features from **multiple layers** in the model	Individual intensity scaling
XGrad-CAM[122]	Confidence-weighted gradients: $W_{l,k,c} = \frac{1}{\lvert i \times j \rvert} \sum_i \sum_j \frac{\text{gradient}_{i,j,l,k,c} \cdot \text{feature}_{i,j,l,k,c}}{\frac{1}{\lvert i \rvert} \sum_i \text{gradient}_{i,j,l,k,c}}$	All features from the last feature extraction block	Individual intensity scaling
Smooth Grad-CAM++[124]	Noise-smoothed gradients: $W_{l,k,c} = \frac{1}{\lvert i \times j \rvert} \sum_i \sum_j \alpha_{i,j}^{k,c} \cdot \text{ReLU}\left(\frac{1}{n} \sum_{n=1}^{n} D_1^k\right)$	All features from the last feature extraction block	Individual intensity scaling
Gradient ×Input	Pixel-level gradients of input: $W_{l,k,c} = \text{gradient}_{i,j,l,k,c}$	Input images	Individual intensity scaling
Method (Others)	**Weight definition**	**Feature and layer definition**	**Scaling method**
Score/Ablation/ Shapley-CAM [123, 130]	Perturbation-based weights	All features from the last feature extraction block	Individual intensity scaling
Integrated Score (IS) CAM[126]	Integrated perturbation-based weights	All features from the last feature extraction block	Individual intensity scaling
Smoothed Score (SS) CAM[125]	Noise-smoothed weights (based on Score-CAM)	All features from the last feature extraction block	Individual intensity scaling
Masking	Model score variation from masked regions	Input images	Individual intensity scaling
Eigen CAM[127]	Principal components of feature matrices	All features from the last feature extraction block	Individual intensity scaling
Multi-/Fusion-CAM[128, 129]	Combined gradients + relevance/self-matching weights	All features from **multiple layers** in the model	Individual intensity scaling

The lack of appropriate intensity scaling can lead to inconsistencies between CAMs and the DL model, resulting in misinterpretation of CAMs and models. Fig. 4.2 (a) presents a schematic diagram of the intensity scaling employed in current algorithms and the consequential errors in different image analysis tasks; (b) presents a conceptual solution that builds a global intensity scale to solve the errors. The activation ranges, representing the upper and lower bounds of signal intensities in CAMs given specific inputs, could vary from input to input. Consequently, the individual intensity scaling may cause same activation values projected to different values in the normalized CAMs, and leads to further interpretation problems.

Not being able to distinguish the most important (highly-contributing) features within the selected layer(s) limits a CAM's application to merely a visual inspection tool, as it provides no insight into the features learned over the entire population. Fig. 4.3 presents a few examples that generate CAMs using all features and randomly

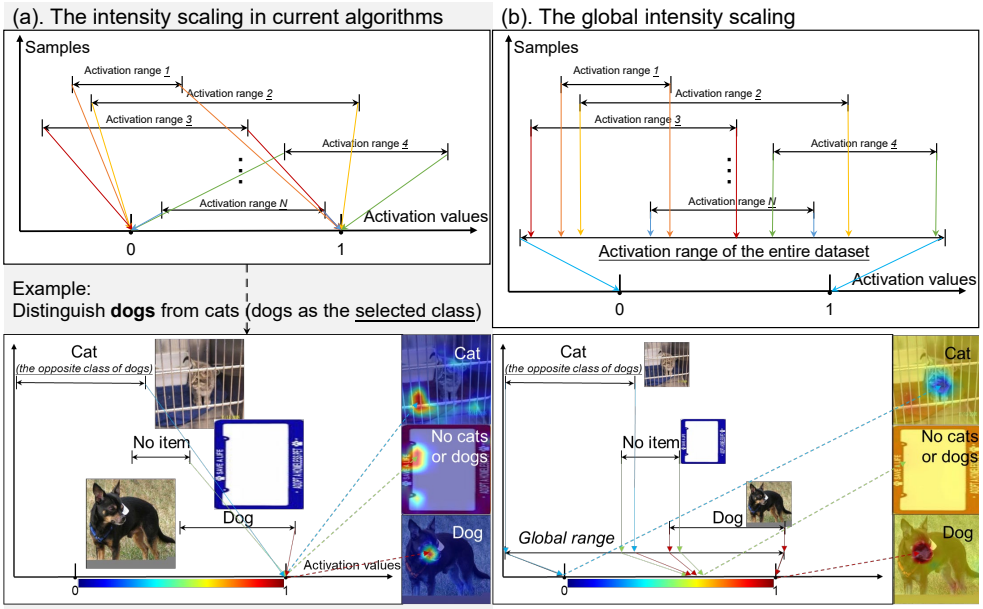


Figure 4.2: (a) A schematic diagram of scaling intensities individually and some examples of the consequential errors. (b) A conceptual solution by global scaling and the results using the same input after solving the errors. In distinguishing dogs from cats, the model needs to look at the cat's face in order to conclude that it is not a dog. However, a completely irrelevant area is highlighted in the example produced by current CAM algorithms based on (a), whereas the globally scaled CAM in (b) is consistent with our understanding of this classification. Similarly, if no cats or dogs are present, there should be no specific focus. The CAM scaled through (a) still highlights some specific areas with activations that are insignificant on a global scale (b). Finally, when the model is presented with an image of a dog, both CAMs in (a) and (b) clearly show focus on the dog's face. *Activation*: the signal intensity of a pixel in an non-normalized CAM. *Activation range*: the activation range of the CAM given a specific input/sample.

chosen 20% of distinct features from the same selected layer(s). As can be seen, using some subgroups of 20% features can generate CAMs that are very similar to using all features, indicating the inequivalence among these features, potential model redundancy at the layer(s) and potential relationship between certain image patterns and feature extraction paths.

To address the above two issues, we propose a framework, which is essentially a feature analysis on a population level. The framework statistically analyzes DL-extracted features within the selected layer(s) at a population level, determining the contribution of each feature within the selected layer(s) during the reasoning, to achieve a global intensity scale and within-layer feature distinction. Since the framework is data-driven

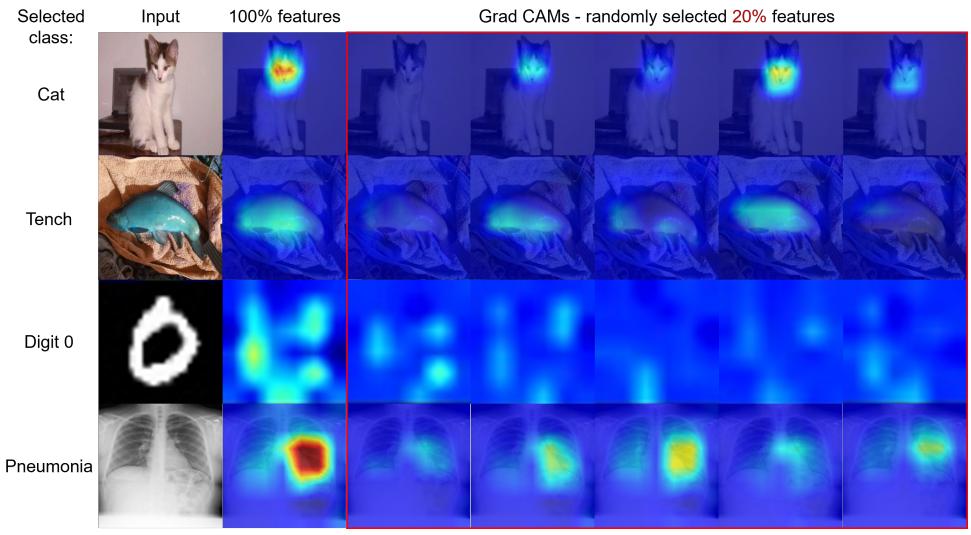


Figure 4.3: Examples of generating CAMs from the same selected layer(s), using all features and randomly selected distinct subgroups of features containing 20% of the total amount of features. The CAMs generated by using randomly selected 20% features differed from each other and especially from the CAMs using all features in the selected layer(s). This phenomenon demonstrates the contribution of some features in the selected layers are not contributing equally to the DL model's reasoning, some features are not even contributing to model's reasoning on some specific output classes.

and relying on very few assumptions, it can be applied to CAM algorithms of various weight definitions and layer selections to improve their performance. Furthermore, the framework can provide insights into other characteristics of the model, such as overfitting, confounders, outliers, model redundancies and the most salient features, which are absent in previous algorithms.

Furthermore, a challenge in evaluating current CAM algorithms is that, most CAM algorithms conduct evaluations based on the intersection over union (IoU) between the focus of CAMs and the semantic objects in images, which has been proven to be biased [134]. Therefore, in addition to only validating through classical IoU-based metrics, we propose another quantitative approach for evaluating the consistency between CAMs and DL models, by calculating the correlation between the mean activations of a CAM and the corresponding logits from the DL model.

We used eight mature datasets from different fields and modalities with accessible and well-trained models to conduct experiments, minimizing the potential errors caused by poorly performing DL models. The experimental results show visually and quantitatively that our framework helps to improve existing CAM algorithms to

overcome the errors, which are caused by a lack of a global intensity scale, and to determine features' importance within the selected layer(s).

We summarized the contribution of this paper as follows:

- We theoretically analyzed the fundamental issues in existing CAM algorithms, and through this process we successfully explained many erroneous phenomena in applying CAMs in practice.
- The proposed feature analysis framework provides an approach to build a global intensity scale for correctly displaying CAMs, and therefore allows correct interpretation of DL models based on these CAMs.
- The feature analysis framework also provides an approach to determine the importance of a feature within the selected layers. This ability enables feature-level analysis for DL models and detection of overfitting, confounders, outliers, model redundancies and the most salient features, extending the use of CAMs.
- We propose a new quantitative approach for evaluating CAMs by calculating the consistency between CAMs and the DL models' outputs. This approach brings in an unbiased evaluation without the need for semantic segmentation, ground truth masks or bounding boxes of any objects in images.

4.2 Method

To address the two limitations, achieving proper intensity scaling and distinguish the difference among features regarding their contribution to the model's outputs, we propose a framework to analyze the activations of each feature within the selected layer(s) across the dataset used for training. Through this process, the statistical information is recorded for establishing a global intensity scale and an "importance" matrix is generated to determine the contribution of each feature at a population level. Fig. 4.4 presents a schematic diagram of the framework.

4.2.1 Global intensity scale

The interpretation problems caused by individual intensity scaling are ubiquitous, yet typically ignored due to confirmation biases and lack of proper evaluation methods [134], especially in CAMs for DL models trained on small datasets. As presented in Fig. 4.2, the solution to this issue is simple – establish a global intensity scale.

Through the process in Fig. 4.4 (a), the global intensity scale could be easily established using the maximum and minimum, which are recorded during the pre-CAM calculation phase, across the training set, and then applied to the validation

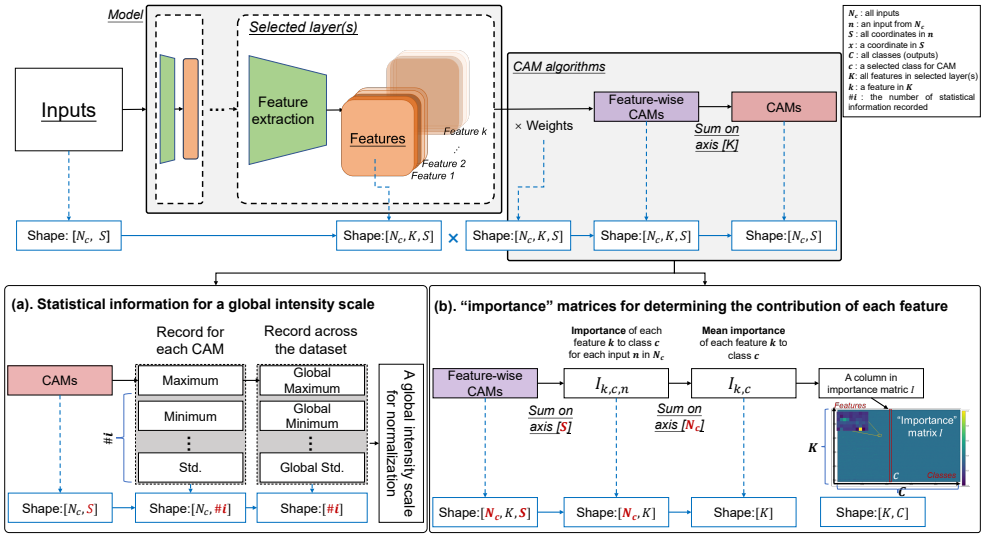


Figure 4.4: The proposed framework for establishing a global intensity scale and distinguishing the "importance" of different features within the selected layer(s). (a) indicates the process of establishing a global intensity scale for CAM normalization, through recording simple statistical information from CAMs at a population level. Based on this analysis, the global maximum and minimum or global percentiles are used for performing normalization. (b) presents the process for calculating contribution (importance $I_{k,c}$) of each feature k to the selected class c regarding the CAMs for each feature. The process starts from an intermediate stage that already exists but not explicitly named in current CAM algorithms - Feature-wise CAMs. These feature-wise CAMs are the products of CAM-defined weights and feature $k \in K$ within the selected layer(s), before summing along the axis of K. Instead of summing along axis K, the "importance" $I_{k,c}$ of each feature $k \in K$ is calculated by summing along axes of spatial coordinates and inputs.

set or new datasets. Furthermore, the activations below zero are also preserved and considered as part the minimum during this process while in current algorithms they were typically discarded.

Regarding the scaling method, we choose the hyperbolic tangent function as the intensity scaling projection function, because some outliers with extremely high or low activations in some instances could lead to a considerable dispersion of maximum and minimum. This dispersion may further lead to some displaying and interpretation problems, as some regions could be overly emphasized or underrepresented. The intensity scaling $\sigma()$ is then defined by:

$$\sigma_{P_{\text{low}}}^{P_{\text{high}}}(x) = \tanh\left(\alpha \cdot x - \beta\right), \tag{4.1}$$

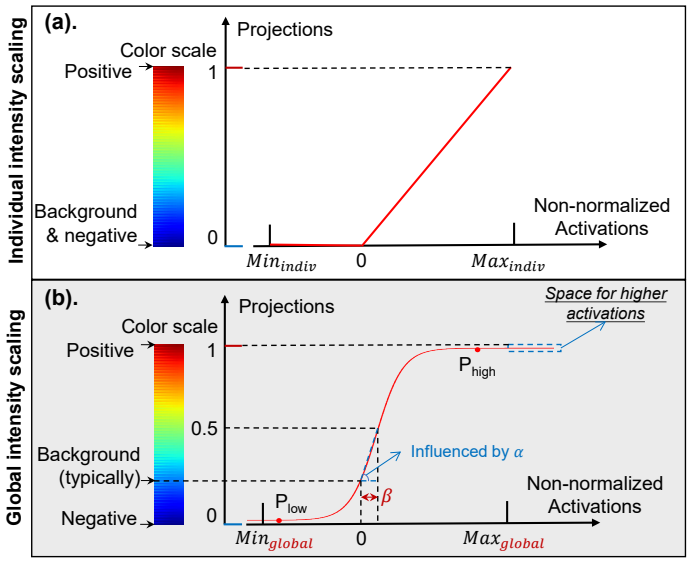


Figure 4.5: Illustration of the scaling methods. (a) The intensity scaling in current algorithms, discarding the activations below zero and projecting based on individual maximum and minimum; (b) The intensity scaling using hyperbolic tangent function to project the activations to displayed CAMs, preserving the negative activations. The proposed intensity scaling method has some advantages, including: (1) the infinite domain of definition that enables accurate CAMs on other datasets with higher/lower activations than the current maximum and minimum; (2) the ability to control the projection curve for better CAM display; and (3) preserved negative activations that can provide more insights into model's decision. Without applying the global intensity scaling, discarding activations lower than zero typically results in a fixed activation of zero (lowest in the CAMs) for the background. When there are no activations above zero in the non-normalized CAMs, discarding these activations lower than zero could even lead to highly-activated backgrounds as shown in the introduction section. After applying the global intensity scaling, the lowest activations come from the opposite-class objects, which would vary from dataset to dataset and model to model. Consequently, the activation level of the background varies in a range from light blue to yellow (JET color map) as the projected value of zero activation shift according to β.

$$\alpha = \frac{\tanh^{-1} V_{\text{high}} - \tanh^{-1} V_{\text{low}}}{P_{\text{high}} - P_{\text{low}}}, \tag{4.2}$$

$$\beta = \frac{P_{\text{high}} \cdot \tanh^{-1} V_{\text{high}} - P_{\text{low}} \cdot \tanh^{-1} V_{\text{low}}}{P_{\text{high}} - P_{\text{low}}}. \tag{4.3}$$

where α and β are constants, calculated through P_{high} and P_{low}. P_{high} and P_{low} represent the values of V_{high}th and V_{low}th percentiles regarding the activations in CAMs. α controls the steepness or slope of the curve and β shifts the curve horizontally.

As shown in Fig. 4.5, this function enables a non-linear projection that its domain of definition allows the activations above/below the maximum/minimum of the dataset used for calculate the global intensity scale. This space is reserved for extremely strong signals that only appear in new data, representing new characteristics which should be considered.

As the proposed intensity scaling and preservation of negative signals are essentially a functional correction to existing CAM algorithms, we preferred to not call it a new CAM method, but rather an intensity-scaled version of these methods, labeled with subscript "gs" meaning "globally scaled" . For example, for Grad CAM, in the rest of text, "Grad CAM_{gs}" represents CAMs with a global intensity scale, while "Grad CAM_{is}" or "Grad CAM" is used to indicate CAMs with the original intensity scale.

4.2.2 Feature distinction

Similar to the idea of CAMs where high values represent higher "importance" or more attention, we define the "importance" of a feature (a particular forward path) to a specific class c as the mean activation of this feature k to class c across the entire dataset. This "importance" , represented by $I_{k,c}$ (feature k and selected class c), is given by Eq. 4.4.

$$I_{k,c} = \frac{1}{|N_c|}\sum_{n\in N_c}\left(\frac{1}{|S|}\sum_{x\in S}(W_{k,c,n}(x)\cdot F_{k,c,n}(x))\right) \tag{4.4}$$

Where $I_{k,c}$ is the importance for feature k that belongs to K, the set of all features in the selected layer(s), and the selected class c from all output classes C; $n \in N_c$ refers to an input case in N_c, where N_c is the set of samples with a size $|N_c|$; x, a coordinate, represents a voxel or pixel from the set of coordinates (S) from the inputs; $W_{k,c,n}(x)$ and $F_{k,c,n}$ are the weight and feature value at the coordinate x based on feature k, class c and input n. In some CAM algorithms like Grad CAM, the $W_{k,c,n}(x)$ can be simplified to $W_{k,c,n}$ because they have a fixed weight for all coordinates in feature k.

Fig. 4.4 (b) presents an illustration of this simple calculation, starting from a feature-wise CAM that already exists in current algorithms. By calculating the mean activation within the feature k, the general contribution (the activations of all the pixels/voxels) within feature k is projected into a single value $I_{k,c,n}$ for each input n. Accumulating this single value across the entire dataset, the general contribution of this feature k is then recorded and compressed from a series of values into a mean value.

This mean value $I_{k,c}$ is then calculated for all different selected classes $c \in C$ and different features $k \in K$ to formulate an "importance" matrix I, of which the horizontal axis represents all the classes C and the vertical axis refers to features K. The pseudocode can be seen in Alg. 1

Algorithm 1 Computation of Feature Importance Matrix I

Input:

- C: Set of all output classes.
- K: Set of features in selected layer(s).
- N_c: Samples of class $c \in C$ (size $|N_c|$).
- S: Spatial coordinates (pixels/voxels) in input (size $|S|$).
- $W_{k,c,n}(x)$: Weight for feature k, class c, sample n at x.
- $F_{k,c,n}(x)$: Activation of feature k for sample n (class c) at x.

Output:

- I: Importance matrix $|K| \times |C|$, where $I_{k,c}$ is the importance of feature k to class c.

Initialize : $I \leftarrow$ Zero matrix of size $|K| \times |C|$
for *each class* $c \in C$ **do**
  for *each feature* $k \in K$ **do**
    sum_importance $\leftarrow 0$ **for** *each sample* $n \in N_c$ **do**
      spatial_sum $\leftarrow 0$ **for** *each coordinate* $x \in S$ **do**
        spatial_sum $\leftarrow$ spatial_sum $+ W_{k,c,n}(x) \cdot F_{k,c,n}(x)$
      mean_activation $\leftarrow$ spatial_sum$/|S|$ sum_importance $\leftarrow$ sum_importance $+$ mean_activation
    $I_{k,c} \leftarrow$ sum_importance$/|N_c|$
return I

Through defining importance matrix above, the most informative features in the selected layer(s) can be then found and selected to generate CAMs with these features alone. Similarly to the CAM_{gs}, which refers to the CAM with a global intensity scale, we use CAM_{sf} (CAM using "selected features") to refer to CAMs that are generated using selected features in the following context.

4.2.3 Evaluation: global intensity scale

Evaluation in previous CAM algorithms typically consists of two methods - qualitative evaluation by visual inspection of the CAM and quantitative evaluation by object localization [59]. The visual inspection is performed to ensure that the highlighted regions in the CAMs are consistent with the foci of human knowledge in classifying the inputs. The quantitative evaluation measures how much the confidence is affected by masking the regions in the image that are highlighted in the CAMs, using the so-called "average increase" and "average decrease/drop" . The original image and the image masked by the CAMs (thresholded and normalized) are fed into the model. This produces two output confidence values (one from the original image and one from the masked image).The average increase and decrease are then defined as the difference between the two confidence values.

However, these evaluation metrics have some substantial drawbacks: the changes in these metrics can be caused by both the model and the CAM algorithms, but they are

only used to evaluate the CAMs, without excluding the impact of model performance and human knowledge involvement. For example, CAMs that always highlighted many regions in an input, because the model was not well-trained and always "looked" everywhere, could also receive consistent average increases and decreases in the quantitative evaluation based on object localization. Furthermore, the evaluation of CAMs by the accuracy of target object localization relies on certain assumptions: (1) a perfect model that has perfect performance and obtains output classes "only" based on the target objects - without any inference from the environment or other objects that might correlate with the output classes. This assumption is so strict that no model has yet been shown to satisfy it. (2) The CAM algorithms must have no prior knowledge that would lead to good object localization without DL models - a randomly initialized model without training could get "convincing" CAMs to pass visual checks by an over-designed CAM algorithm [134].

Derived from the very fundamental idea of CAMs – an input with important information for classification should be more highlighted in the attention maps than those with irrelevant information – we propose a new metric that measures the consistency between CAMs and the model's logits. This avoids introducing human knowledge of the "meaningful" objects and is therefore independent of the performance of a DL model. The metric in this study, defined as the correlation coefficients Rs between the sum of class activations and the model's logits, focuses on the relationship between CAMs and the model's outputs. The use of logits instead of confidences may improve readability, as the activation functions (e.g. the SoftMax function) break the linear relationship between logits and confidences by rescaling according to the logits for other classes, leading to a non-linear relationship between model logits and class activations. The formal definition of this metric is given by:

$$r = \frac{\sum(L_i - \bar{L})(A_{i,c} - \bar{A}_c)}{\sqrt{\sum(L_i - \bar{L})^2}\sqrt{\sum(A_{i,c} - \bar{A}_c)^2}}, \tag{4.5}$$

$$\bar{L} = \frac{1}{n}\sum L_i, \quad \bar{A}_c = \frac{1}{n}\sum A_{i,c} \tag{4.6}$$

Where L_i represents logits of the DL model given a specific input i, A_i refers to the sum of activations in the CAMs for a particular output c given the same input i, n is the number of samples in the dataset for evaluation.

4.2.4 Evaluation: feature distinction

Feature distinction based on the importance matrix is proposed to differentiate between the importance of features within the selected layer(s). In order to validate whether important features are indeed contributing most, we calculated the correlation between the mean intensity of the original CAM and the model's logits using, and tested whether this correlation remained when using only 5-25% most important features to

calculate the CAM. If the correlation does not drop considerably, then these selected features are indeed the most important.

Therefore, we generated the CAMs using the 5-25% most important features and calculated the same correlation-based metrics as the CAMs using all features based on Eq. 4.5. By comparing the metric changes, we can quantitatively validate the reliability of the proposed feature distinction.

Similarly, if the top 5-25% of features in the importance matrix are indeed important, the model outputs should remain the same if the most important features are retained and the irrelevant features are filtered out. Since feature masking does not alter the intensity distribution of the inputs, and therefore does not significantly affect the performance of the model, we propose a quantitative evaluation metric based on feature masking. Specifically, we propose to block the forward paths of the unselected features during inference, to obtain the logits based on the selected features only, and then take the accuracy changes as the evaluation metric. The formal definition is given by:

$$\Delta_{acc} = Acc(f) - Acc(f_{mask}) \tag{4.7}$$

Where f represents the DL model, f_{masked} refers to the model with a mask for specific features (neglecting some forward paths in the model) and Acc is the accuracy. The higher Δ_{acc}, the more important these masked features are.

The prevailing metrics of average increase ($Avg.Inc$) and decrease/drop ($Avg.Dec$) in previous studies studies [59] were also applied with the same idea to validate the effectiveness of feature distinction - if the top 5-25% of features in the importance matrix are indeed important, the $Avg.Inc$ and the $Avg.Dec$ should stay the same or not decrease substantially. The definitions of these metrics [59] are given by:

$$\Delta_{AvgInc} = AvgInc(f) - AvgInc(f_{mask}) \tag{4.8}$$

$$\Delta_{AvgDec} = AvgDec(f) - AvgDec(f_{mask}) \tag{4.9}$$

Where $AvgInc$ and $AvgDec$ (average drop/decrease) are defined [59] as:

$$AvgInc(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(f(\hat{x}_i) > f(x_i)) \times 100\% \tag{4.10}$$

$$AvgDec(f) = \frac{1}{N}\sum_{i=1}^{N} \frac{\max(f(x_i) - f(\hat{x}_i), 0)}{f(x_i)} \times 100\% \tag{4.11}$$

Where i denotes the index of the input sample, ranging from 1 to N. N represents the total number of test samples. x_i refers to the original input image corresponding to sample i. $f(x_i)$ represents the model's predicted confidence score for the target class using the original input image x_i. $\hat{x}_i$ refers to the modified input image, where important regions have been masked based on the saliency map produced by the

CAMs. $f(\hat{x}_i)$ denotes the model' s predicted confidence score for the target class using the modified image $\hat{x}_i$. $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if $f(\hat{x}_i) > f(x_i)$ and 0 otherwise. $\max(a,0)$ ensures that the computed drop in confidence does not become negative, keeping only positive differences.

Similarly, the Area Under the Curve (AUC) metrics for Insertion and Deletion [59] follow a similar formulation to evaluate the reliability of feature importance maps. These metrics compute the integrated change in model confidence as top-ranked features are incrementally inserted or removed. Specifically, if the saliency map correctly identifies important regions, inserting them should quickly increase the model's confidence (high AUC), and removing them should rapidly decrease the confidence (low AUC). The AUC scores are defined as follows:

$$\text{AUC}_{\text{Ins/Del}} = \sum_{t=1}^{T-1} \frac{f(I_t) + f(I_{t+1})}{2} \cdot (x_{t+1} - x_t) \tag{4.12}$$

where t denotes the step index ranging from 1 to T, and T is the total number of steps. x_t represents the proportion of features modified (either inserted or deleted) at step t, normalized within the interval $[0,1]$. I_t is the intermediate image at step t, created by either inserting or deleting the top-x_t fraction of features based on the saliency map. $f(I_t)$ denotes the model's predicted confidence score for the target class given image I_t. These AUC values are computed using the trapezoidal rule for numerical integration.

4.3 Materials

We selected eight datasets from different domains, which have been thoroughly explored in previous studies, as the experimental materials to validate the robustness and generalizability. Some recent public datasets were excluded mainly due to the lack of high-performance and publicly-available trained models or the high complexity of well-performing models in these datasets (making CAM visualization difficult). Likewise, we selected eight different CAM algorithms as baseline methods and discarded some recent algorithms because of the following reasons: (1) the source code was not provided, (2) the algorithm was slow and therefore rarely used by other studies, and (3) the algorithm fits into one of the calculation processes in Fig. 4.1 that have already been applied and therefore cannot contribute to further prove the effectiveness of the proposed method.

4.3.1 Datasets and models

The datasets and corresponding models used for method validation include:

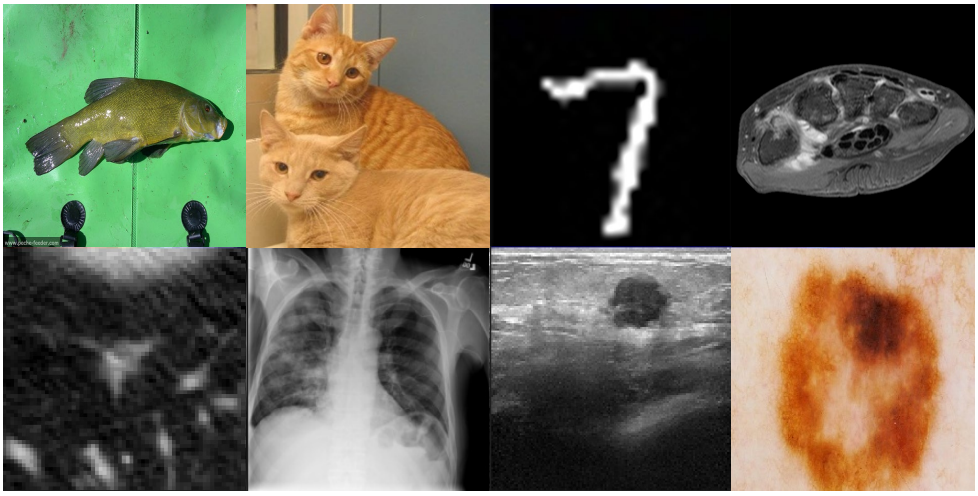


Figure 4.6: Some examples from the image datasets used. Images from the top left to the bottom right are: a natural image from 1000 classes, a cat image from two-class Cats&Dogs, a digit, an MRI scan, a CT scan, an X-ray, an ultrasound image and a skin photograph.

- ILSVRC2012 [135] for 1000-class natural objects, using standard ResNet and VGG as models, with top-1 accuracies ranging from 70% to 83%;
- Cats & Dogs [136] for classifying cats and dogs, using standard ResNet and VGG as the models, achieving accuracies around 99.5% (ResNet34) and 98.6% (VGG16) on 10000 test images;
- MNIST [137] for classifying ten-class digits, using a model consisting of a simple two-layer multi-head self-attention block (from Transformers) with a multi-layer perceptron (fully connected layers), achieving an accuracy of over 99.9% on test digits;
- MRI (T1-weighted contrast-enhanced with fat suppression) from the ESMIRA project [138] for classification of rheumatoid arthritis (RA). This (non-public) dataset, containing over 6000 3D MRI scans from 2000 subjects, includes three classes: early arthritis patients (EAC), patient with clinically suspect arthralgia (CSA) and healthy controls (ATL). The task is to discriminate between these classes. The model used is a 2D plus 3D U-net encoder with a multilayer perceptron that achieved an AUC of 83%;
- Cropped CT scans from the LIDC/IDRI malignancy detection database [139], where the target objects are the malignant lesions. A copy of VGG11 is applied and received an AUC of 81%;

- X-rays from the RSNA pneumonia detection task [140]; the target objects are the regions with signs of pneumonia. The standard ResNet34 received an AUC of 84% on this dataset;

- Ultrasound for breast cancer classification [141], and the target objects are the cancer-related regions. A simple multilayer CNN described in the literature [141] was reproduced and achieved an AUC of 76%;

- Skin images from the SIIM-ISIC melanoma classification [142]; the target objects are the regions associated with malignant skin cancers. The model for this task achieved an AUC of 78% and is a copy of VGG11 (resulting in low resolution due to down-sampling).

4.3.2 CAM algorithms

The CAM algorithms involved in the experiment are Grad CAM [58], Grad CAM++ [59], xGrad CAM [122], Score CAM [123], Smooth Grad CAM++ [124], SS-CAM [125], IS-CAM [126] and pixel-wise Grad CAM (pixel-wise gradients with features). The quantitative evaluation were mainly applied to Grad CAM [58], Grad CAM++ [59], xGrad CAM [122] and pixel-wise Grad CAM, as these methods are the most widely-used in application-oriented studies (e.g. DL applications in medical images) due to having high calculation efficiency.

For the reasons of not using other CAM algorithms, please see the discussion section.

4.4 Experiments and results

Three sets of CAMs were generated on all eight datasets, including the original CAMs based on the existing CAM algorithms, the corresponding CAMs with a global intensity scale, and the CAMs generated using 5-25% most important features (for different datasets, respectively) and the global intensity scale.

- The first subsection focuses on intensity scaling – to validate the effectiveness of the proposed global intensity scale, the correlations with the models' logits were calculated for the original CAMs and the CAMs with a global intensity scale using Eq. 4.5.

- The second subsection focuses on feature distinction, three different evaluation metrics were calculated for validating the effectiveness of the proposed feature distinction method. The first evaluation is the comparison of CAMs using top 5-25% features with the CAMs using all features regarding consistency with DL model's logits (using Eq. 4.5), especially after applying the global intensity scaling. The second evaluation compares the model's performance before

and after masking the specific groups of features (using Eq. 4.7) for different datasets, respectively. The last evaluation follows the calculation of average increase, decrease, insertion and deletion and then compares the difference between these metrics while masking the unselected features, using Eq. 4.8, Eq. 4.9, Eq. 4.12.

- The third subsection provides some visual examples to show that the global intensity scale helps to improve the explainability and the top 5-25% (consistent with the feature distinction) features can generate very similar CAMs to using all features.

The 5-25% percentages of top features in feature distinction are determined by manually looking at the average activations of each feature in the importance matrices and arbitrarily picking the features whose activation values are substantially higher than the others. There could be some potential approaches to find the optimal percentages, but the definition of "important" features is fuzzy, making it difficult to evaluate the choices. Since the selection of the exact values of these percentages is an open question, we just arbitrarily chose these percentages for simple illustration of the feature distinction's capability according to visual checks on the importance matrices.

4.4.1 Global intensity scaling

Fig. 4.7 presents the scatter plots that show the relation between the sum of class activations in CAMs and the model' s logits. In the figure, the columns with odd indexes are the quantitative results based on original CAM algorithm, and the columns with even indexes are the results based on the proposed global intensity scaling based on the algorithm. Table 4.2 provides more numerical details including the correlation coefficients and average increase in the coefficients after applying the global intensity scaling, comparing the original CAM algorithms and those with the global intensity scale.

In Fig. 4.7, most CAMs received high (nearly 100%) correlation coefficients using the global intensity scale (especially for pixel-wise Grad CAM$_{gs}$ and XGrad CAM$_{gs}$), however, some CAM algorithms may perform worse than others. These errors may originate from the weight definition, for example, for the scatter plots of MNIST-Grad CAM and MNIST-Grad CAM$_{gs}$, the calculation of the weights of Grad CAM is based on the average of the gradients in the feature, which is not appropriate for the MNIST dataset, where the digits appear more frequently in the center of the image.

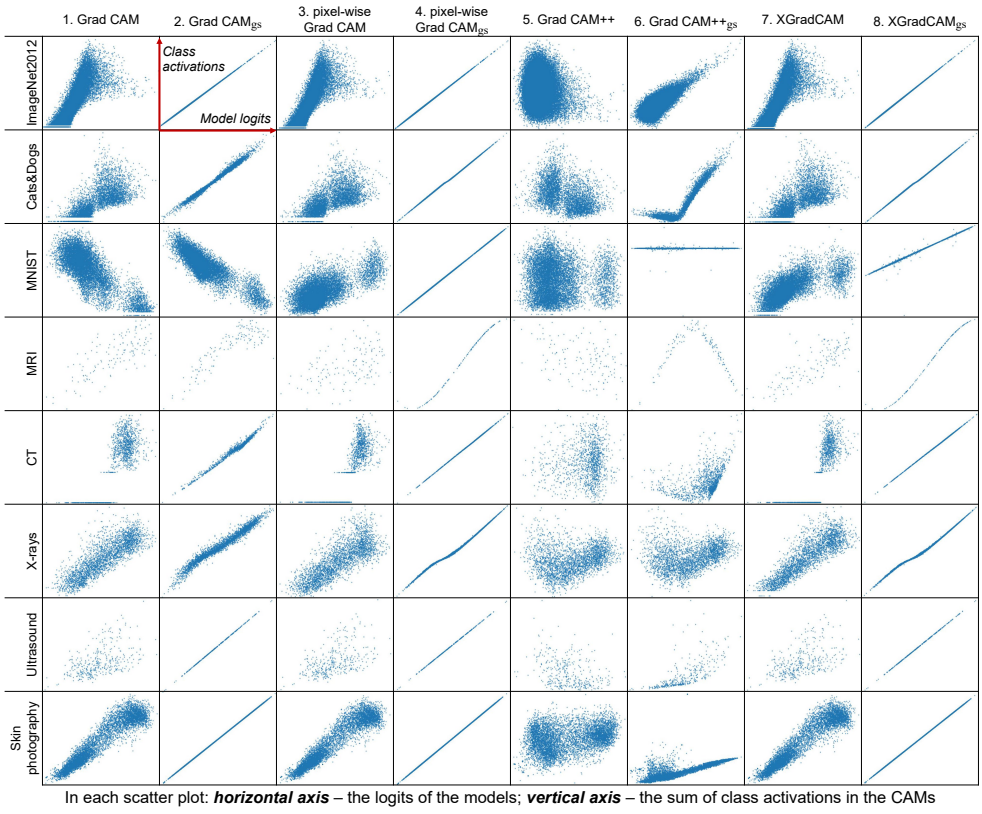


Figure 4.7: Scatter plots for the sum of class activations in CAMs (vertical axis) and the logits of the models (horizontal axis) in each dataset, comparing the CAMs based on existing methods with current individual intensity scale and the ones with the global intensity scale. In most cases, the global intensity scaling (marked with $_{gs}$) significantly improved the correlation between CAMs and model outputs.

4.4.2 Feature distinction using importance matrix

Top 5% (top 25% ImageNet and top 20% for MNIST) features, based on the average class activations of each feature in importance matrices, were selected as the most important features and used for the following evaluation of the proposed feature distinction.

Using the first evaluation method, Fig. 4.8 and Table 4.3 presents the scatter plots and numerical results on the consistency with DL model's logits, in which the results using all features and the global intensity scaling (namely CAM_{gs}, refers to CAMs with global intensity scale) were compared with CAMs using selected features and the global intensity scale (namely $CAM_{gs,sf}$). In the figure, the columns with odd indexes

Table 4.2: The Pearson's correlation coefficients (r) with the logits of the models, based on the original CAM algorithms and the CAMs using a global intensity scale. All results are statistically significant with p values less than 0.005.

r	Grad CAM	Grad CAM_{gs}	pixel-wise Grad CAM	pixel-wise Grad CAM_{gs}	Grad CAM++	Grad $CAM++_{gs}$	XGradCAM	$XGradCAM_{gs}$	Mean Δ_r(%)
ImageNet2012	0.892	0.999	0.891	0.999	0.048	0.651	0.892	0.999	23.1%
Cats&Dogs	0.754	0.995	0.686	0.999	-0.422	0.820	0.729	0.999	51.6%
MNIST	-0.678	-0.811	0.545	0.999	0.019	0.029	0.708	0.998	15.5%
MRI	0.704	0.763	0.457	0.999	-0.331	-0.271	0.752	0.999	22.7%
CT	0.553	0.980	0.758	0.999	0.071	0.485	0.765	0.999	32.9%
X-rays	0.849	0.983	0.827	0.999	0.292	0.292	0.874	0.999	10.7%
Ultrasound	0.532	1.000	0.532	1.000	-0.340	0.825	0.532	1.000	64.2%
Skin photography	0.946	0.999	0.945	0.999	0.266	0.790	0.945	0.999	17.1%

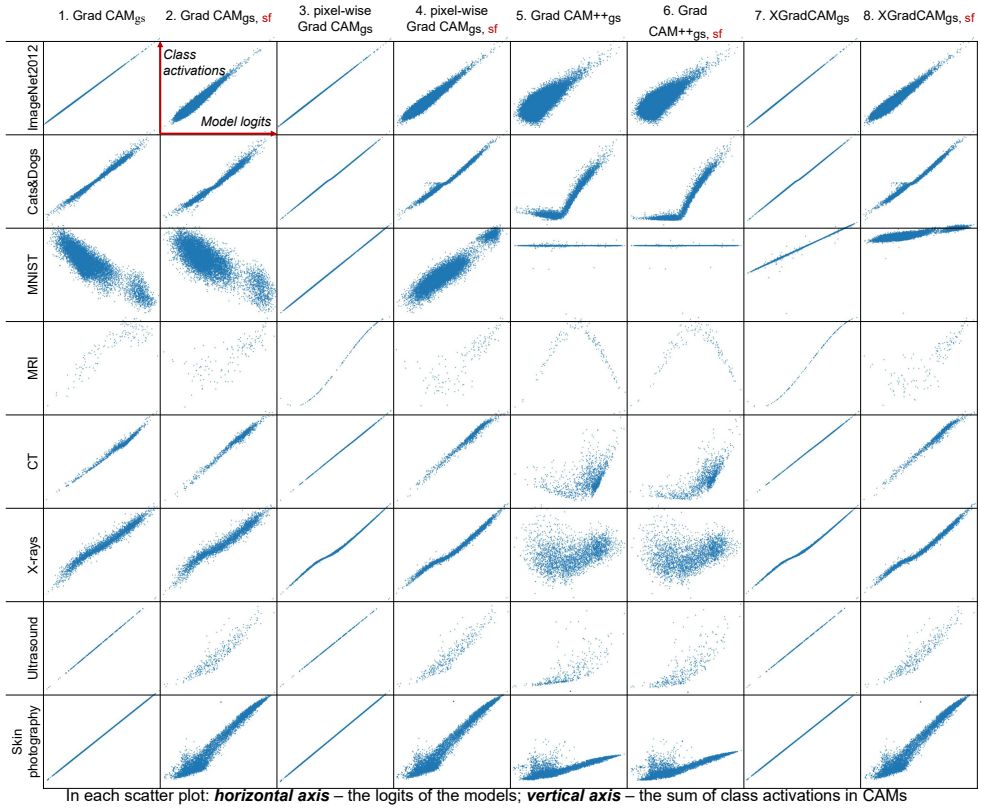


Figure 4.8: The scatter plots for the sum of class activations in CAMs and the logits of the models in each dataset, comparing the CAMs based on all features available at the selected layers of the models and on principal features (top 5% for six datasets, 20% for MNIST and 25% for ImageNet2012). Although excluding most features leads to losses in consistency, the class activations of the CAMs based only on principal features maintained a high level of consistency with the logits of the models.

Table 4.3: The Pearson's correlation coefficients (r) of the CAMs and the logits of the models, comparing the CAMs with a global intensity scale before and after applying the feature distinction to select the most important features. Ave.Dec represents the average decrease of the correlation between the activation values in CAMs and models' logits.

r	Grad CAM$_{gs}$	Grad CAM$_{gs,sf}$	pixel-wise Grad CAM$_{gs}$	pixel-wise Grad CAM$_{gs,sf}$	Grad CAM++$_{gs}$	Grad CAM++$_{gs,sf}$	XGradCAM$_{gs}$	XGradCAM$_{gs,sf}$	Mean Δ_r (%)
ImageNet2012	0.999	0.915	0.999	0.914	0.651	0.652	0.999	0.915	6.35%
Cats&Dogs	0.995	0.992	0.999	0.993	0.820	0.871	0.999	0.992	-0.88%
MNIST	-0.811	-0.703	0.999	0.728	0.029	0.049	0.998	0.728	11.33%
MRI	0.763	0.797	0.999	0.849	-0.271	-0.143	0.999	0.856	3.28%
CT	0.980	0.972	0.999	0.981	0.485	0.752	0.999	0.981	-5.57%
X-rays	0.983	0.978	0.999	0.995	0.292	0.355	0.999	0.996	-1.27%
Ultrasound	1.000	0.927	1.000	0.927	0.825	0.829	1.000	0.927	5.37%
Skin photography	0.999	0.931	0.999	0.931	0.790	0.815	0.999	0.931	4.47%

Table 4.4: The accuracies and Δ_{acc} (Eq. 4.7) of the models inferring on eight datasets, based on top 5–25% features and all features available at the selected layers of the models.

Dataset	Acc. (All features)	Acc. (Keep 5–25% selected features)	$\Delta_{acc}\uparrow$ (Masking 5–25% selected features)	Acc. (Unselected features)	$\Delta_{acc}\downarrow$ (Masking unselected features)
ImageNet2012 (Natural image)	0.706	0.689	0.672	0.034	0.017
Cats&Dogs (Natural image)	0.985	0.962	0.478	0.507	0.023
MNIST (Digit)	0.994	0.933	0.919	0.075	0.061
ESMIRA (MRI)	0.831	0.772	0.299	0.532	0.059
CT (cropped LIDC/IDRI)	0.813	0.772	0.289	0.524	0.041
X-rays (RSNA pneumonia)	0.839	0.797	0.333	0.506	0.042
Ultrasound (usbc)	0.764	0.642	0.253	0.511	0.122
Skin photography (SIIM-ISIC)	0.780	0.735	0.255	0.525	0.045

are the quantitative results based on CAM_{gs}, and the columns with even indexes are the results after applying feature distinction to retain the top 5-25% features and drop others.

In all these CAM algorithms and datasets, the CAMs, generated by the features selected through the proposed feature distinction, show a very close level of consistency with the model's logits. These results demonstrate that the features selected by the proposed feature distinction indeed retained the most important information since it can keep the consistency with the model's logits, with a small decrease in correlation coefficients (see the last column in the Table 4.3).

Subsequently, using the second evaluation, the DL models' performance before and after masking some specific features was compared to explore how much these features can influence the models' accuracies. The results in Table 4.4 presents the accuracies while different feature groups were masked. These models' accuracies decreased only slightly if the top 5-25% features (according to the proposed method) were retained during models' reasoning, and dropped considerably when these "important" features were masked and others are retained. These results demonstrate that these selected "important" features are indeed most influential to the models' outputs.

Likewise, in the third evaluation, the average increases and decreases [59]of the original CAMs and the CAMs generated through selected features were compared to

explore how much these features can affect these models' outputs. As shown in the results in Table 4.5, the changes of the average increases and decreases, by comparing the original CAMs and the CAMs using selected features (CAM_{sf}), demonstrate the highlighted regions in CAM using selected features could cover most of the important regions in the original features. These results further prove that these selected "important" features are indeed influential to models' outputs.

Table 4.5: The changes of the average increases and decreases (drops), and the changes of AUCs for insertion and deletion based on different CAM algorithms, comparing the original CAMs and CAM$_{sf}$s. The last column named Average represents the average increase/decrease of the CAM$_{sf}$s comparing to the original CAMs. For average increase and insertion, the higher the better; for average decrease and deletion, the lower the better. Percentage changes in metrics were applied to increase readability.

Dataset	ΔMetric	GradCAM	Pixel-wise GradCAM	GradCAM++	XGradCAM	Average
ILSVRC2012	AvgInc (↑)	-6.01%	-11.16%	1.85%	-4.20%	-4.88%
	AvgDec (↓)	2.67%	6.44%	-1.50%	2.04%	2.41%
	Insertion (↑)	-5.28%	-0.13%	-1.33%	-4.26%	-2.75%
	Deletion (↓)	7.10%	1.78%	3.23%	5.45%	4.39%
Cats&Dogs	AvgInc (↑)	-6.22%	-10.68%	-7.41%	-11.02%	-8.83%
	AvgDec (↓)	9.09%	23.81%	8.33%	19.05%	15.07%
	Insertion (↑)	-0.07%	-0.36%	-0.02%	-0.30%	-0.19%
	Deletion (↓)	1.93%	4.32%	1.26%	3.21%	2.68%
MNIST	AvgInc (↑)	100.00%	-33.33%	0.00%	0.00%	16.67%
	AvgDec (↓)	-24.02%	0.55%	12.62%	11.49%	0.16%
	Insertion (↑)	0.23%	-0.61%	-0.80%	0.56%	-0.16%
	Deletion (↓)	-0.25%	0.65%	1.09%	-0.45%	0.26%
MRI	AvgInc (↑)	-1.85%	-4.07%	-1.57%	-1.47%	-2.24%
	AvgDec (↓)	-1.42%	1.79%	4.66%	-2.16%	0.72%
	Insertion (↑)	-7.90%	-4.54%	-0.27%	-1.54%	-3.56%
	Deletion (↓)	7.84%	3.36%	0.35%	3.75%	3.83%
CT	AvgInc (↑)	-5.73%	0.37%	-5.92%	-3.52%	-3.70%
	AvgDec (↓)	10.20%	1.83%	4.83%	10.83%	6.92%
	Insertion (↑)	0.31%	-0.84%	-0.05%	-1.07%	-0.41%
	Deletion (↓)	-0.85%	1.79%	1.36%	1.41%	0.93%
X-ray	AvgInc (↑)	-0.79%	-7.14%	10.43%	-7.32%	-1.20%
	AvgDec (↓)	1.20%	8.88%	-1.20%	8.45%	4.33%
	Insertion (↑)	0.00%	1.00%	0.45%	-0.14%	0.33%
	Deletion (↓)	0.12%	1.07%	-0.04%	1.06%	0.55%
Ultrasound	AvgInc (↑)	117.95%	0.00%	-61.54%	-12.90%	10.88%
	AvgDec (↓)	1.67%	-5.21%	0.00%	4.10%	0.14%
	Insertion (↑)	0.28%	-0.56%	3.28%	0.43%	0.86%
	Deletion (↓)	-0.02%	0.83%	-3.12%	-0.57%	-0.72%
Skin Photo	AvgInc (↑)	-2.87%	-2.87%	-0.63%	-2.87%	-2.31%
	AvgDec (↓)	0.79%	0.79%	0.80%	0.79%	0.79%
	Insertion (↑)	-0.89%	-0.88%	0.44%	-0.89%	-0.56%
	Deletion (↓)	0.13%	0.18%	-0.20%	0.28%	0.10%

4.4.3 Visual examples

Before the visual examples, the difference in the color scale before and after applying the proposed intensity scaling needs to be clarified, as they have a different reference as the lowest-activated objects. In the original intensity scaling, the background activations or zero activation object in the non-normalized CAMs are treated as the reference. Under this setting, the lowest activations represent no contribution. After applying the global intensity scales, all activations are retained, and the lowest activations now represent the negative contribution. This shift in lowest-activation reference from background to the opposite objects results in a green or light blue background in the CAMs after applying the global intensity scale, and reversely highlighted areas (blue instead of red) that represent objects that indicate the other classes. The color scale on the left sides of Fig. 4.5 (a) and (b) present an illustration on this difference.

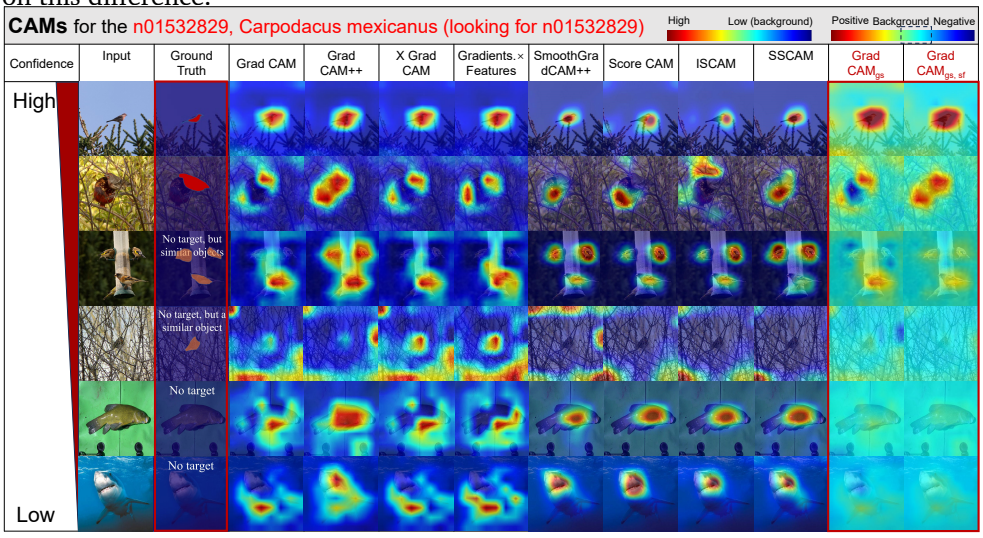


Figure 4.9: CAMs for a selected class of Carpodacus mexicanus, which are based on different CAM algorithms and the same ResNet34 model weights. With the decrease in confidences from the model, the CAMs are supposed to have decreasing class activation levels, yet the trends only appear in the Grad CAM$_{gs}$ and the Grad CAM$_{gs,sf}$.

Fig. 4.9 to 4.11 briefly show some examples based on existing class activation mapping algorithms compared to the CAMs with the global intensity scaling based on the P_{90} and P_{10} (P_{high} and P_{low} in Eq. 4.5) from the distribution analysis. In the visual checks, we present the Grad CAM$_{gs}$ and Grad CAM$_{gs,sf}$, which is based on the most basic CAM algorithm, to show the importance and effectiveness of the global intensity

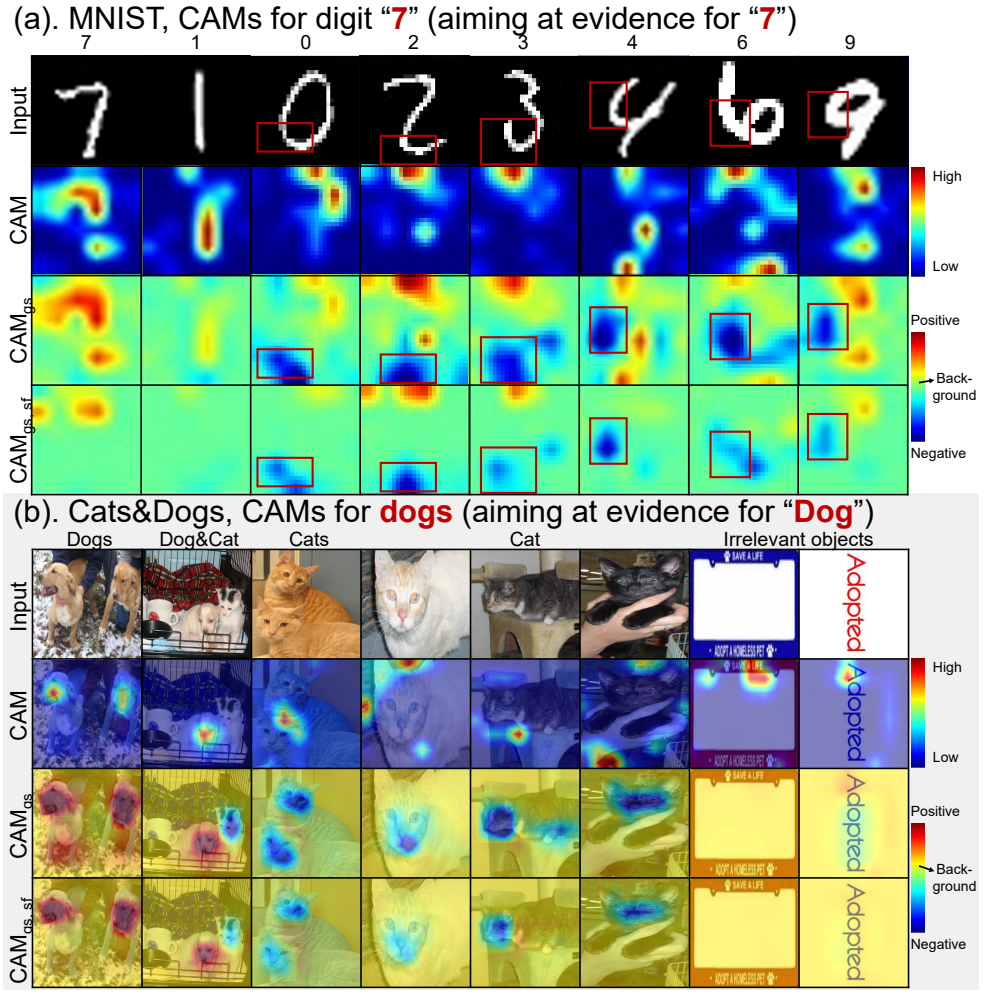


Figure 4.10: Examples of CAMs for (a) MNIST (with a selected class of digit "7"), (b) Cats & Dogs (with a selected class of "dogs"). In every row, the images are ordered as: input image, Grad CAMs, Grad CAM$_{gs}$s, and the scaled CAMs with feature selection (Grad CAM$_{gs,sf}$).

scaling and the results that it could exceed other CAM algorithms on matching the model confidences.

Fig. 4.9 presents an example of the CAMs for a selected class of Carpodacus mexicanus (ImageNet2012) based on a standard ResNet34. From the perspective of the consistency with the confidences, Grad CAM$_{gs}$ and Grad CAM$_{gs,sf}$ with global intensity scaling managed to show a decreasing trend on their CAMs as the confidences decreased, while other methods failed.

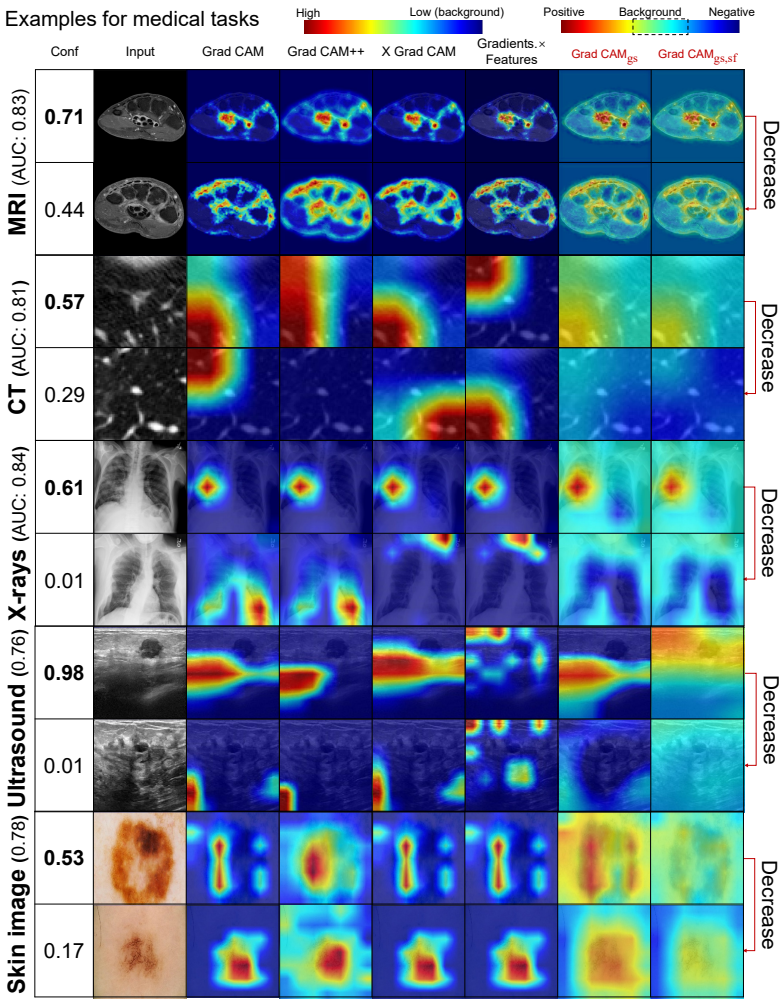


Figure 4.11: Examples of CAMs for medical images with different confidences based on models. Please note that some saliency maps highlight regions without any visible objects or lesions and the models give output classes of diseases or lesions. This could happen for different reasons: (1) the models were not trained well; (2) the models found some confounders that were not visible; (3) the models found the images were "not normal enough" and overall tended to treat it as patients and put all those were not normal enough to patients; (4) the visualization method went wrong. In this figure, the Grad CAM$_{gs}$s and Grad CAM$_{gs,sf}$s show better consistency with the confidences, while CAMs based on other CAM algorithms get more class activations from the input with lower confidences than the higher confidences, especially regarding the influence of background.

Fig. 4.10 presents an example of the CAMs for a selected class of digit "7" [137]

based on a model with a two-layer multi-head self-attention block followed by a multilayer perceptron, and an example of the CAMs for a selected class of "dogs" [136] based on a standard VGG16. In the CAMs for digit "7" , the red boxes indicate the regions that have negative contributions to the model' s decision to give an output class of "7" according to Grad CAM_{gs}. This results fits our intuition since there should be no bright pixels around the left and bottom left regions around the digit "7" and are supposed to have some intensities appearing on the top regions. In the CAMs for the selected class "dogs" , the Grad CAM_{gs} did not only remove the random highlighted regions in the background when the input was a cat or an undefined object. The Grad CAM_{gs} gave class activation levels of less than the background for cats as they are the opposite class of dogs in this dataset, and gave overall "background" class activation levels to the undefined inputs.

Fig. 4.11 presents some examples of the CAMs with a selected class of lesion existence, inflammation areas or any signs of diseases (the opposite classes of normal/healthy) based on the five different modalities in medical imaging, using the models described in the Material section. These medical examples prove, that the same problem of inconsistency between class activations and model logits also exists in the original CAMs. In these applications, scalar outputs are more frequently required than in the previous datasets. In the medical domain scalar outputs (logits or confidences) should demonstrate the severity of diseases/symptoms or the chances of a developing disease. With a global intensity scale, medical images with higher model confidences on lesions, inflammation areas or diseases received more and higher-class activations than those with lower model confidences.

4.5 Discussion

In this paper, we proposed a framework of feature analysis to establish a global intensity scale and distinguish the importance of different features, based on the same premise as the CAM algorithms. It is not a new CAM algorithm or a new definition of weights for CAMs, but a different perspective of using existing weighted features. The intensity scaling and feature selection could be applied to most CAM algorithms without increasing computation time during inference, serving as a useful functional supplement, to minimize the risk of interpretations based on wrong CAMs. The feature distinction succeeds in distinguishing the importance of features within the selected layer(s) and find the most important features for a specific class. In the following subsections, we discuss about the uses of the proposed feature distinction based on the importance matrix, generalizability of the proposed method and the definition of importance. Some other topics (e.g. regression tasks, time costs and evaluation metrics) can be found in the supplementary materials.

4.5.1 Uses of the feature distinction

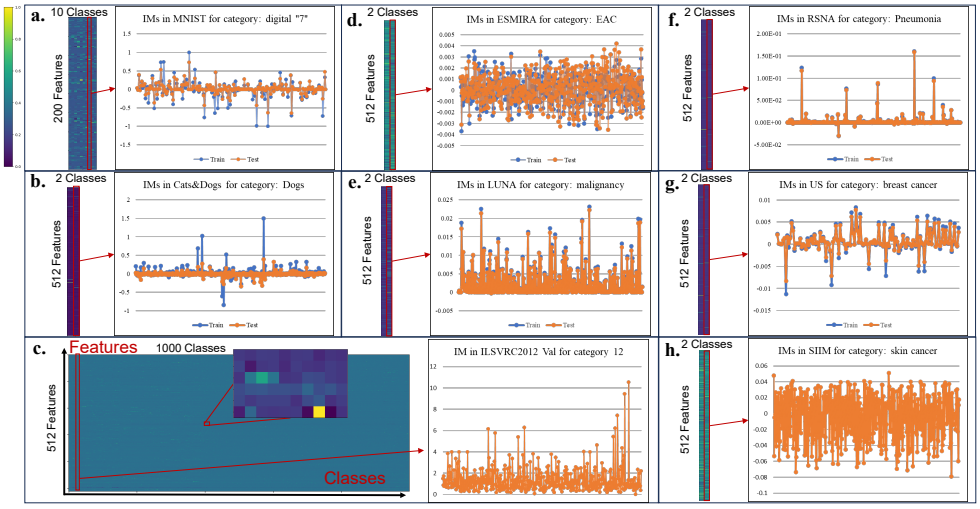


Figure 4.12: The importance matrices (resized for better visualization) in feature distinction for models trained on: (a) MNIST, (b) Cats & Dogs, (c) ILSVRC2012 validation set only, (d) ESMIRA dataset where the target classes are EAC, (e) Lung Nodule Analysis 2016 (LUNA) for malignancy detection (based on lesions), (f) RSNA Pneumonia Detection Challenge for pneumonia existence, (g) breast ultrasound images dataset, and (h) SIIM-ISIC melanoma classification, targeting the melanomas. In each group, the entire importance matrix (IMs) is shown on the left and a detailed line graph of the column (red boxes in the IMs) is on the right, which is targeting a particular class and then subtracting the mean activation level in the IMs. The values represent the overall importance of certain features to the model decision, the line graphs of importance matrices based on the training set are shown in blue, and the ones based on the test set are shown in orange.

The importance matrix for feature distinction could provide useful information about both the model and dataset, such as:

- **Overfitting**. The difference in importance matrices between the training and test sets indicates potential overfitting when the class activations of each feature in the training set differ considerably from those in the test set. For example, in the line graphs of Fig. 4.11 (a), (b), (d), (e), (f) and (g) the importance matrices are shown from both the training and test sets. From the difference between the training and test sets, (a) and (b) indicate more severe overfitting, as they have features that are much more activated in the training sets than in the test sets, while the training and test sets share a very similar pattern in other line graphs.
- **Principal features**. If we select a single column in the importance matrix (thereby selecting a class), each value represents the average class activation

level of a specific feature for that selected class. Those features that are more activated (higher class activations) also have a greater influence on this selected model output class in general. Therefore, they can serve as the main features of this study. For example, in the line graphs of Fig. 4.11 some features are more frequently activated (except for 11.d and 11.h), resulting in a large ratio of total activation values to average importance over the entire dataset. The values of these features indicate the potential of some features to emerge from many others.

- **Potential confounders**. Confounders in deep learning models usually refer to some objects or noise in the background of images or time series that could provide a "shortcut" for a model to be easily trained and make inferences based only on these non-targets. In the importance matrices, confounders could cause extremely high values for some specific features and relatively low values for all other features because they provide the "shortcut" . These potential confounders can therefore be checked by generating CAMs using only the relevant features. As the models are copies of successful models on these public datasets, we did not find any significant signs of confounding in these datasets. In the line graph of Fig. 4.11 (f) the model may benefit from some potential confounders compared to other datasets, some visual checks on the CAMs are valuable for further investigation.

- **Special cases and outliers**. Special cases or outliers would show very different class activations of the features compared to the importance matrices of their corresponding classes. For example, in the blue boxes within the line graph of Fig. 4.3, we present a simple example from (d) of the ESMIRA project, where the sample belongs to the class "EAC" , but has a different picture of activated features and was classified as the opposite with confidences around 0.5.

- **Model redundancies**. The ratio of highly activated features out of all features in the importance matrices represents the ratio of influential features in the model, as the values are a combination of activation frequency and "amplitude" . Therefore, the rarely activated features in all classes/outputs contribute very little to the model decision in most cases and could be removed by model pruning. For example, in the line graphs of Fig. 4.11 (b) and (f), compared to other importance matrices, they have more features that are rarely activated across the datasets, while some features are significantly activated. This phenomenon indicates that the models have some "lazy" features that contribute very little and can be pruned.

4.5.2 Definition of the importance

For the definition of "importance" in this paper, we took the average of the sum of class activations over the whole dataset as the "importance" of a feature to a specific class. This originates from the same idea behind CAMs: the more activated features contribute more to the final output and are therefore the more important features. However, the definition of the importance of each feature could be further investigated, as the standard deviation of the average of the weighted features could also be informative. For examples, a rarely activated feature with massive class activation values in only very few cases might obtain high values in the importance matrices based on current definition of "importance" . However, this feature may represent a serious confounder or artifacts that only appear in some cases of a certain class because of unknown reasons. Considering this kind of situation, using average of the weighted features can lead to misunderstanding in some cases. Considering this, it is difficult to find an optimal way of defining the feature importance, delivering simple and clear messages. The definition of feature importance is therefore still an open question, awaiting the exploration through more studies.

4.5.3 Why are some CAM algorithms not included in the experiments?

Some CAM algorithms were not presented in the experiments and results. There are three main reasons: (1) They fell in the scope of the formula in Fig. 4.1, and therefore are theoretically the same as the used CAM algorithms. (2) The source code of some CAM algorithms were not accessible. (3) The computational efficiency was an essential issue, especially for perturbation-based CAM algorithms. In most studies, Grad CAM and Grad CAM++ are always the first choice just because of better technical support and high computing efficiency. (4) For multi-layer CAM algorithms, the low-level features have no direct link to the model outputs without going to the high-level features, their information was covered by the high-level features unless there are skip-connections to the output. Moreover, the calculation of weights becomes unreliable because of the increasing non-linearity of models during back-propagation or backward calculation. The advantage of these multi-layer methods is the resolution of CAMs, which is another topic out of the scope of this paper.

4.5.4 Metrics like Dice or IoU to evaluate CAM algorithms?

Metrics such as Dice and IoU with masks of ground truths introduce an assumption that models trained through classification or regression loss functions should naturally focus on the region of interests that is aligned with human knowledge to obtain the output. However, models may predict through confounders, artifacts and even other objects in the foreground that have a high correlation with the targets. Since CAM algorithms are mainly developed to interpret trained models, not to cater to

confirmation bias, these metrics based on the assumption were not applied.

4.5.5 Generalizability of global intensity scaling

The framework is generalizable to other datasets, without needing to recalculate the upper and lower limits, provided that they result in similarly distributed class activations. This was accomplished by applying a non-linear projection for the global intensity scaling, that provides some redundant space for displaying extreme values. If a new dataset produces a different distribution of class activations, the framework only needs the inference process to update the upper and lower limit for rescaling all CAMs, without needing to retrain the DL model. Therefore, the time cost of the framework is proportional to the DL model' s inference costs, which is considerably less than for training. While the inference process is still the limiting factor of the framework, the following strategies in this study were designed to improve this. Since the goal of global intensity scaling is to make an unbiased comparison between CAMs within the same dataset, inference on the entire training dataset is not always necessary. P_{high} and P_{low} could be calculated on the test set, a subgroup of the training set with random sampling or even a subgroup of the test set. This approach has been applied in the experiments on the ILSVRC2012 validation set. If the purpose is to have correct and accurate visualization of only a limited amount of inputs, the P_{high} and P_{low} can also be computed on those inputs only.

4.5.6 Generalizability of the feature analysis

Although the proposed methods are mostly applied to convolutional neural networks, the importance matrix and the common intensity scaling are not limited to these type of networks. They are also feasible for other deep learning models, because the original CAM algorithms are also feasible for different types of models [130]. We had applied the proposed method to the tiny transformer on MNIST to prove its generalizability. However, the integrated feature analysis faced a transition problem that requires further studies on the definition of features in transformers and recurrent neural networks and the spatial reorganization to generate CAMs in these models. ,

4.5.7 Hyper-parameter issues

The proposed method introduces several hyper-parameters to the original CAM algorithms, including the parameters for color re-scaling and threshold of "top N%" features. These hyper-parameters may significantly influence the visualization and therefore requires proper design. The 5-25% percentages of top features in our experiments are manually and arbitrarily determined, which may not be optimal. This open question about the approach to define the "principal features" and obtain the "top N%" requires further investigation and may vary from dataset to dataset.

4.5.8 Other topics

For reproducibility and reused of the method, the code for the method, evaluation metrics and links to datasets can be found in `https://github.com/YanliLi27/IFA`. Furthermore, for better and clearer message delivery, we and put some important experiments and discussion in the supplementary materials of [57]. In the supplementary materials, we present a detailed discussion about: (1) the proper evaluation for model interpretation methods, the impact of confirmation bias and subjectivity in some of the visual examples and current evaluation metrics. (2) An alternative evaluation metric for the consistency in classification tasks, with a similar probabilistic interpretation of the area under the receiver operating characteristic (AUROC) in two-class classification. (3) The reason of using logits instead of confidence (after Softmax) to evaluate the correlation. (4) Generalizability of this method on regression models. (5) Time costs of the proposed method and potential acceleration.

4.6 Conclusion

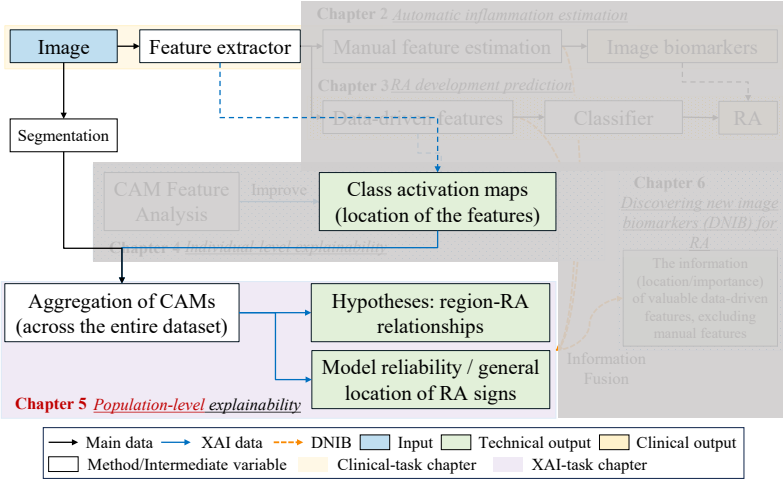
In summary, we propose a feature analysis framework, which establishes a global intensity scale and determines the importance of features, to fix the errors in current CAM algorithms and extend uses of CAMs. It helps to improve the consistency between CAMs and models, and provide more insight in both models and datasets. As the framework relies on very few assumptions, and considering the potential uses of the feature distinction in analyzing models and datasets, the proposed framework is expected to avoid errors caused by intensity scaling, extend the uses of CAM algorithms and facilitate model interpretation.

Acknowledgment

This work is supported by the Netherlands Organization for Scientific Research (NWO, TTW 13329), the European Union's Horizon 2020 research and innovation programme (No.714312) and the China Scholarship Council (No.202108510012).

5

Aggregation of Class Activation Maps for Explaining Deep Learning at a Population Level



This chapter was adapted from:

Yanli Li, Xikai Tang, Denis P. Shamonin, Hessam Sokooti, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Johan H.C. Reiber, and Berend C. Stoel. "Aggregation of Class Activation Maps for Explaining Deep Learning at a Population Level." (*submitted*)

Abstract

Class activation maps (CAMs) are frequently used to help understand the inference by deep learning (DL) models, especially in medical imaging that requires high reliability and explainability. If CAMs consistently contain high activations in regions with specific image patterns that experts also focus on during their clinical decision making, the reliability of the model is made plausible. However, since a single CAM provides only a qualitative illustration from one case, aggregation of CAMs is required to generally validate DL models on a population-level. This aggregation involves determining the correspondence between anatomical/pathological regions in the original images among all subjects in a population. Subsequently, the class activation in these corresponding regions can be compared to the local image patterns, that are known to significantly contribute to the final classification (i.e. regions containing particular image patterns with high importance). Currently, this aggregation is performed manually in observer studies, which are time-consuming and potentially biased. Therefore, we propose an automatic CAM aggregation workflow, based on image segmentation. To validate the aggregation method objectively, without the need for an observer study, ground truth data is needed on the importance of specific image patterns for indicating a particular disease. We defined this importance as the posterior probability of a disease, given the presence of particular 'pathological regions'. Therefore, a series of simulations was conducted, in which different classification tasks, in which different posterior probabilities of a virtual disease given specific image patterns were defined. We tested whether the mean activation in a particular pathological region changed consistently with an induced change in the ground truths of the predefined posterior probabilities. The results of the simulated classification tasks demonstrated the reliability and sensitivity of the proposed CAM aggregation, with a high correlation of 0.98 between the known posterior probabilities of 'pathological regions' and the mean activations in these regions in the CAMs. The results also demonstrate that the CAM aggregation can generate hypotheses about the causality between these regions and the clinical outcomes. Finally, the validated aggregation method was applied to two medical datasets from different modalities (MRI and X-rays), to prove its generalizability and potential applications in clinical practice. We expect that the proposed CAM aggregation can help validate neural networks at a population level and reveal possible correlations between regions containing specific image patterns and clinical outcomes.

5.1 Introduction

Deep learning (DL) methods, which enable automatic data-driven feature engineering and decision-making, have achieved strong performance across various image-related applications, including applications in medical imaging. Before DL models can be widely deployed on a large scale, an essential question needs to be addressed: why should DL models be trusted? Good performance alone does not guarantee trustworthiness, as DL models can make decisions based on confounding factors rather than real informative image patterns. This concern is particularly substantial in risk-averse domains [105, 106, 107], like medical image analysis, where decision-making accuracy is paramount.

To mitigate the risk and validate the reliability of DL models, the most intuitive solution, alongside the use of even larger-scale data for validation, is to improve the explainability of DL models and ensure the alignment with expert knowledge. For image processing, this validation refers to ensuring that the focus of DL models in images is consistent with expert knowledge. This consistency is typically validated by studies, in which human observers visually check attention maps of a DL model and investigate whether the highlighted regions align with expert knowledge and contribute meaningfully to the model's output. The most widely-used techniques for generating attention maps for DL models are class activation maps (CAMs) and their variants [55, 58, 59], in which the signal intensity is assigned to each region based on its contribution to the outcome. CAMs with a global intensity scale have been shown to achieve high agreement with DL model outputs [57], enabling the representation of DL models' focus and facilitating the validation of DL models by manually checking, whether CAMs have highlighted the regions that are involved in the experts' inference processes.

However, human observers of a DL model must repeat the visual checks over the entire dataset, as a single CAM provides only a qualitative illustration for a particular input and cannot represent the entire dataset. This tedious work must be done each time before implementing new DL models, a time investment that cannot be ignored, especially with large datasets. In addition, model validation based on visual inspection is susceptible to confirmation bias caused by prior distrust or overconfidence in DL models. Therefore, an automatic, unbiased, population-level analysis method is needed to validate and explain DL models.

However, due to the difficulty of aggregating CAMs for developing such an automatic population-level analysis and of validating the analysis method, only few studies have investigated the approach of aggregating CAMs for a population-level conclusion. Cherepanov et al. [143] proposed to aggregate CAMs across the whole population to analyze the population-level contribution of model-extracted features,

but focused on features and models rather than the original image regions. In the medical imaging field, where potential image biomarkers are of interest, Park et al. [144] proposed to register and then overlay CAMs to find out, which image regions in the chest CT contribute more to the estimation of lung function. These studies provided the path of aggregation through registration and is therefore highly dependent on the accessibility and quality of image registration. As a consequence, datasets without high-quality registration need other solutions to achieve similar population-level analysis.

In this study, we proposed an aggregation framework to explain DL models at a population level and quantitatively measure their reliability, using semantic segmentation that are typically accessible and well-investigated in medical image analysis to establish the correspondence of regions in original images and CAMs and aggregate CAMs for conclusions. This framework calculates and assigns to each region an 'importance' -value at a population level, which is defined to be consistent with the posterior probability of a disease (the ground truth class) given a particular image pattern (the presence of an anatomical or pathological region). In test statistics, the posterior probability is equivalent to the positive/negative predictive value of a medical test, which is in our case an image pattern in a particular region. This automatic process can replace tedious repetition in observer studies, by directly comparing the 'importance' -values from the framework and posterior probabilities from expert knowledge of different regions. Furthermore, the framework's 'importance' -value could generate hypotheses regarding the estimation on the posterior probabilities of a class given specific patterns when expert knowledge is unavailable in new domains. We also propose a validation framework for this aggregation method to prove its reliability.

The layout of this paper is as follows. First, we introduce the workflow of the proposed method, including generating and rescaling CAMs, obtaining the pixel-wise locations of the regions through segmentation and combining the abovementioned information to achieve CAM aggregation. To subsequently validate the proposed method, we propose a method that can generate a series of simulation experiments with simulated diseases (essentially virtual classes in medical field), containing different posterior probabilities of the simulated diseases given the presence of particular lesions as a simple region-pattern combination ("RPC"). In the next section, we introduce the simulation datasets and two other medical imaging tasks of rheumatoid arthritis (RA) prediction and cardiovascular stenosis percentage prediction to demonstrate the clinical and technical impact of this method. Thereafter, we present the results and the quantitative analysis after applying the method to the simulated experiments, followed by the analysis on the other tasks. Finally, the limitations, advantages and potential improvements of the proposed method are discussed and

Table 5.1: The glossary for terms used in this study.

Terms	Meaning in this study
Region	An anatomical or pathological area/volume in images
Image pattern	A specific image feature (e.g., presence of a structure, textures, high signal intensities.)
Class	The output of the dataset (e.g., diagnoses of a disease, the classification of an image)
Activation	The signal intensity in the class activation maps
Posterior probability	The posterior probability of a particular class given a specific image pattern in a particular region
Importance	How predictive a region is to a particular class according to DL models
RPC (region-pattern combination)	The combination of an image pattern, occurring in a region, considered as a feature for classification.
Simulated disease / Virtual class	A class created based on particular posterior probabilities of a specific class given specific RPCs for validation

summarized in the last two chapters. Table 5.1 presents a glossary for the terms used in this study.

5.2 Method

Following the framework in Fig. 5.1 (b), the aggregation of CAM consist of four parts – a revised CAM algorithm, a method for region definition, the calculation process of aggregation and a series of simulation experiments that generate classification tasks with different posterior probabilities for validation.

5.2.1 Class activation mapping with a global intensity scale

The existing methods of generating CAMs for deep learning models follow a process with three major steps: (1) weight calculation (based on gradients [58, 59], perturbation weights [123] or other ideas [127]), (2) feature map selection (last convolution layers [122, 58, 123] or multiple layers [121, 128, 129]) and (3) an individual intensity scaling to normalize and visualize CAMs.

In the third step of the above process, however, CAMs are scaled individually using the maximum and minimum of each individual CAM and the class activations below zero are discarded. As a consequence, the signal intensities in the displayed CAMs cannot be compared among different cases [57] (see Fig. 5.1 (a)). Therefore, global intensity scaling in Fig. 5.1 (b) is required for normalizing CAMs to enable cross-case analysis and draw general conclusions at a population-level.

We selected a method, based on which CAMs achieved highest correlations with DL models' outputs in our previous study [57], to build such a global intensity scale for all the CAMs. First, this method feeds the DL model with a group of all cases and collect the distribution information (e.g. maximum, minimum and percentiles) during the CAM calculation. Then, the method uses the maximum and minimum or percentiles to build a global scale. Subsequently, all CAMs are rescaled to this global scale. This proved to increase the correlation between the over activations of CAMs and the corresponding model's outputs.

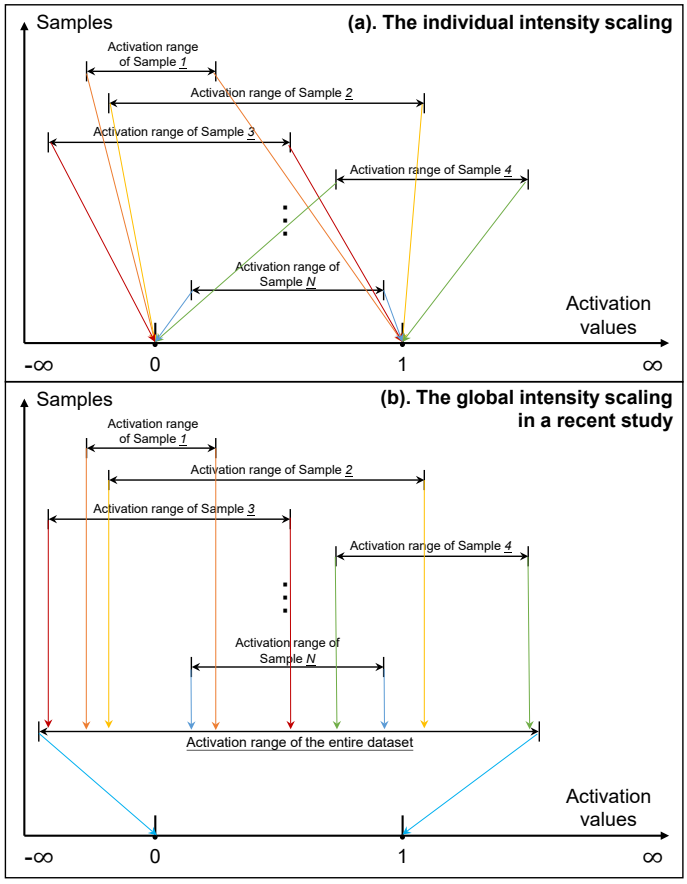


Figure 5.1: The conceptual illustration of (a) individual intensity scaling used in current CAM algorithms and (b) global intensity scaling proposed in a recent study [57].

5.2.2 Region definition through segmentation

Accurate region correspondence among individual cases is the fundamental basis of this aggregation framework. Although the most intuitive solution is to use manual annotations so that the region definition could be as accurate as human experts, manual annotations on large-scale data are however laborious and even more time-consuming than observer studies. An alternative is semantic segmentation, which is a prosperous field in image processing due to the development of fully convolutional networks, especially in medical image processing where U-Nets [78] are widely-used. Semantic segmentation may not be as accurate as manual annotations, but segmentation inaccuracies are predictable, measurable and solvable, by checking the performance of the segmentation models, measuring the errors of specific semantic

classes in the annotated cases and manually fix these errors. With the development of DL methods in the field of medical images, the accessibility of accurate semantic segmentation is foreseeable in many tasks, therefore the CAM aggregation based on the segmentation could also be extended to these tasks.

5.2.3 Aggregation process

The aggregation process aims at obtaining the mean activation of each region obtained by semantic segmentation, and we consider these mean activations as the 'importance'. According to the interpretation and purpose of CAMs, this 'importance' should reflect the overall contribution of patterns in this region to the model's outputs. Therefore, the 'importance'-value of a region intuitively should be consistent with the posterior probability of the disease given those patterns in the region. 5.1 presents the definition and calculation process of the mean activation $M_{i,cl}$ in region i for class cl based on the whole dataset.

$$M_{i,cl} = \frac{1}{|N_i|} \sum_{n \in N_i} \frac{\sum_{x \in C} \left(Seg_{n,i}(x) \wedge A_{n,cl}(x) \cdot \mathbb{I}(Seg_{n,i}(x) > 0) \right)}{\sum_{x \in C} \left(Seg_{n,i}(x) \cdot \mathbb{I}(Seg_{n,i}(x) > 0) \right)}. \tag{5.1}$$

Where N_i refers to a subset of all images, in which the segmentation object i exists, n represents the nth image; $Seg()_{n,i}$ refers to the one-hot segmentation for region i based on the nth input image; $A()_{n,cl}$ represents the calculated CAM for input n and selected class cl; x refers to a coordinate in the coordinate space C inside the nth input image.

The above calculation process typically requires the segmentations and class activation maps to have the same data dimension and shape, as it performs an AND operation at each coordinate x of the CAM and segmentation. However, since some models that down-sample on a specific dimension are also widely applied in medical image analysis, such as 2.5D models, where the segmentations could sometimes be 3D, while the CAMs are 2D. In cases where the segmentation does not match the CAM, a modification can be made to the aggregation process to make an estimation. For example in 2.5D models, the inputs and segmentations have a shape of $[D(depth), H(height), W(width)]$ while the CAMs have a shape of $[1, H, W]$. To multiply segmentations and corresponding CAMs without repeatedly counting some regions in the CAMs, for each segmentation class, we calculate the maximum union of multiple slices of each segmentation class to obtain a new segmentation with a shape of $[1, H, W]$. Then the aggregation process can follow that of the 2D inputs.

5.2.4 The conceptual method for validating the aggregation

As the correctness of our framework is dependent on the accuracy of the aggregation process, ground truths on the posterior probabilities of a class given particular image patterns in a specific region (namely a specific region-pattern combination, RPC) are

necessary for evaluating the accuracy of the aggregation process. For establishing these ground truths from expert knowledge, which is used implicitly by observers during their inference process, needs to be made explicit for a quantitative measurement of the alignment between expert knowledge and DL models.

To make this knowledge explicit and obtain the quantitative measurements, the human expert inference process needs to be formalized. In this formalization, an image can contain different regions, which are defined as anatomical structures or pathological lesions, and within these regions, specific image patterns may occur, such as distinct high signal intensities, textures, or simply the presence of this structure (e.g. the presence of a pathological lesion within an anatomical structure). The combination of image patterns within regions may provide important evidence to a disease. As an actual clinical example from MRI imaging in rheumatology: the combination of degenerative changes at the trapezium of the first carpometacarpal joint (CMC-1), and similar changes at the base of MC-1 bones and absence of other findings, could very likely indicate the class of 'CMC-1 osteoarthritis' [145].

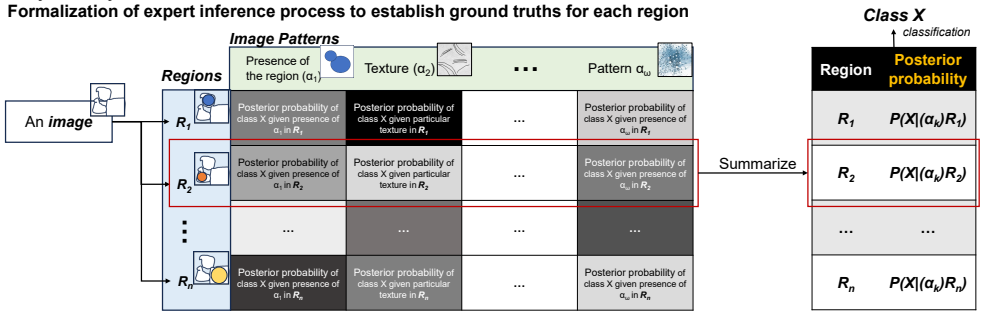


Figure 5.2: A general workflow for the formalization of expert inference process to define the ground truths – the posterior probability of a class given a particular RPC (i.e. the presence of the region, a specific texture in a region). This process provides a ground truth on the contribution of each region to model's outputs.

In this formalized process, human experts obtain the classification by defining anatomical or pathological regions, finding image patterns in these regions, and combining the class-specificities of specific patterns that appeared in particular regions. Three steps are needed to quantify this process and establish the ground truths (of the posterior probabilities) for the aggregation, as presented in Fig. 5.2. The first step is to define anatomical and pathological regions in the original images and the corresponding locations in the CAMs. The simplest solution to automatically define regions is semantic segmentation, as image segmentation is widely investigated in many tasks and is therefore frequently accessible. The second step is to identify image patterns within these regions. For anatomical regions, the image patterns refer

to particular signal intensities, textures and other characteristics. For pathological regions (e.g. lesions), which are frequently regarded as separate segmentation objects in semantic segmentation, the presence or absence of these regions could be seen as an image pattern. Finally, the third step is to quantify the "posterior probability" of a specific class given a particular RPC using statistical models. This posterior probability represents quantified expert knowledge and used as the ground truth for validating the aggregation of CAMs in this study.

The above process provides a general method of creating the ground truths for validating the aggregation, yet the posterior probabilities in real datasets are usually fixed or not accessible in many domains. Given the limitation of real datasets, a simulated dataset, with controllable posterior probabilities of a specific class given particular RPCs, provides better insight into how the 'importance' from DL models vary according to the induced variation in posterior probabilities.

5.2.5 The simulation for validating the aggregation

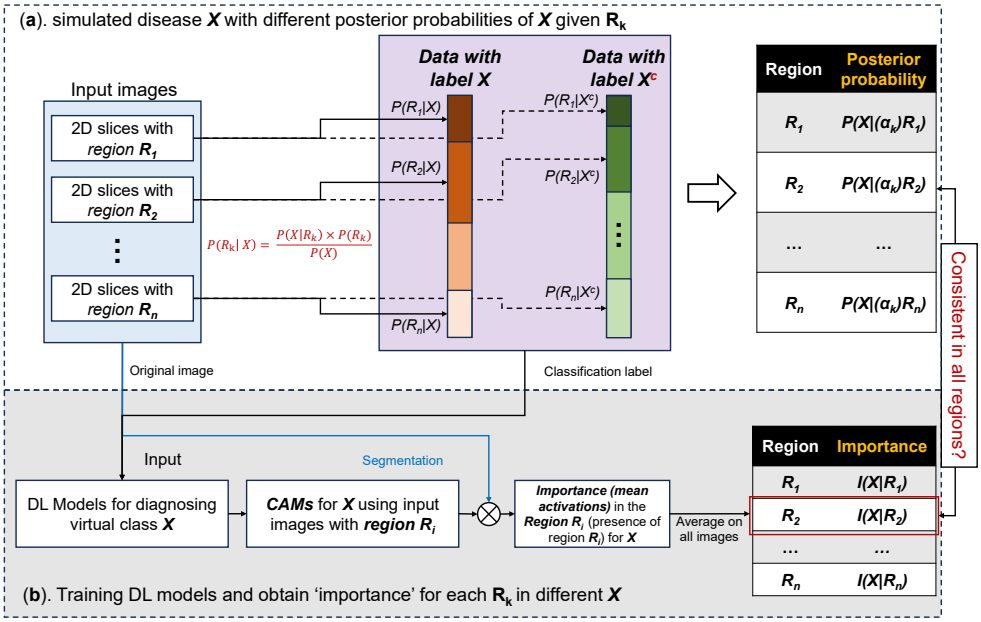


Figure 5.3: Illustration of: (a) the method to simulate a disease X with different and controllable posterior probabilities of X (X^c is the opposite class, namely 'healthy') given disease pattern R_k/ (the presence or absence of region R_k/, such as the presence of lesions). (b) the process of training DL models to diagnose the disease, and then obtaining the importance (defined as mean activations in the region) of each region for comparison with the posterior probabilities to validate DL models. All regions R were defined to fit the mutual exclusion requirement.

To comprehensively validate the proposed CAM aggregation framework, the consistency between the 'importance' -values and the posterior probabilities of a specific class given a particular RPC, needs to be determined under different and controllable posterior probabilities.

However, datasets with classification/regression outputs, well-located regions and controllable ground truths on posterior probabilities of a specific class given particular RPCs are rarely accessible.

We therefore propose to establish a series of simulation experiments that have accurate segmentation and controllable posterior probabilities of a specific class given particular RPCs, based on the following ideas: (1) datasets with manual annotations can provide accurate definitions of regions preventing segmentation errors; (2) Virtual classes(diseases) that have well-designed posterior probabilities of the diseases given some RPCs in the images can be generated by **proportionately assigning images with and without these RPCs** into the group of images with label of 'disease' and 'healthy' ; (3) The presence or absence of a specific region, such as lesions, can be considered a simplified RPC and used to generate a posterior probability of a virtual class given the presence or absence of this region; (4) for 3D images, their 2D slices containing different segmentation objects can serve as the input to increase the number of samples, and minimize potential model errors from the DL models for distinguishing a "virtual disease" from healthy.

Fig. 5.3 presents a general illustration of generating the simulation experiments and validating the CAM aggregation. In this process, we applied a reversed Bayes' Theorem to calculate the proportion of images with specific RPCs in 'disease (X)' and 'healthy (X^c)' , as shown in the reversed Bayes' Theorem Eq. 5.2, the calculation process for the proportion of α_k in the data group of X using predefined posterior probabilities $P(X|\alpha_k)$.

$$P(\alpha_k|X = \frac{P(X|\alpha_k)P(\alpha_k)}{P(X)} \tag{5.2}$$

Where RPC α_k refers to the kth RPC, virtual class X as the disease (opposite to healthy), $P(X \mid \alpha_k)$ is the controllable pre-designed posterior probability for RPC α_k and ground truths for CAM aggregation, $P(\alpha_k \mid X)$ defines the likelihood of observing α_k (exclusive) given disease X, essentially the proportion of images with RPC α_k in the group of disease X. The marginal probability $P(\alpha_k)$ of α_k represents the ratio of images with RPC α_k in all selected images, and $P(X)$ is the prior probability of disease X. The $P(\alpha_k \mid X)$ is then used as the proportion for assigning images to the two groups ('disease' and 'healthy') in Fig. 5.3.

By changing the posterior probability $P(X \mid \alpha_k)$ for different RPC α, a series of classification tasks with different ground truths of posterior probability can be generated and the model trained on these tasks should show changes in importance in the CAM

aggregation, consistent with the changes in the imposed posterior probabilities.

To simplify the problem, minimize the effect of other factors and fulfill the requirement of mutual exclusion in Bayes' Theorem, we applied the following strategies while generating the simulated classification tasks:

- To minimize the influence of data imbalance in DL model training, the prior probability of disease X in all simulation experiments were set to be 0.5.
- The likelihood of observing α_k given X is fully computed based on Bayes' theorem. However, to accommodate the real-world scenario where multiple RPCs coexist while still satisfying the mutual exclusion of Bayes' theorem – the regions α_k should be independent without overlap, we divided images containing a specific RPC α_k into two major categories: those containing only α_k and those where it coexists with other distinct RPCs, ensuring mutual exclusion in the calculation. The posterior probability is then calculated based on the proportion of images that contain the given RPC in all selected images.
- To make the RPCs not as frequent as the background, all RPCs with a posterior probability greater than 0.5 was set to have the marginal probability below 0.5, except for the background and irrelevant (to X) RPCs used as a reference.

5.3 Datasets

To validate the proposed CAM aggregation through the simulation method, we used the KiTS23 dataset from the 2023 Kidney and Kidney Tumor Segmentation challenge (KiTS23) [146], in which segmentation objects include backgrounds, kidneys, tumors and cysts. In the simulation, the meaning of the objects were ignored, and we treated the presence of kidney, tumor and cyst as three separate RPCs, – and generated "disease" X labels based on different posterior probabilities of X given these RPCs. After this validation, we applied the method to a rheumatoid arthritis prediction task using MRIs and a stenosis prediction task using X-rays to demonstrate its potential applications.

5.3.1 Simulation 1: qualitative evaluation

Following the simulation design in the previous section, we considered the presence of kidneys, tumors and cysts as RPC α_1, α_2 and α_3, respectively. For mutual exclusion, images with both tumors and cysts were considered a separate RPC α_4 during generating the dataset (image-label pairs). Consequently, the posterior probabilities of α_2 and α_3 are actually the posterior probabilities of disease X given the presence of exclusive tumors and cysts, and there is another posterior probability of disease X given both the presence of tumors and cysts to achieve mutual exclusion and keep the

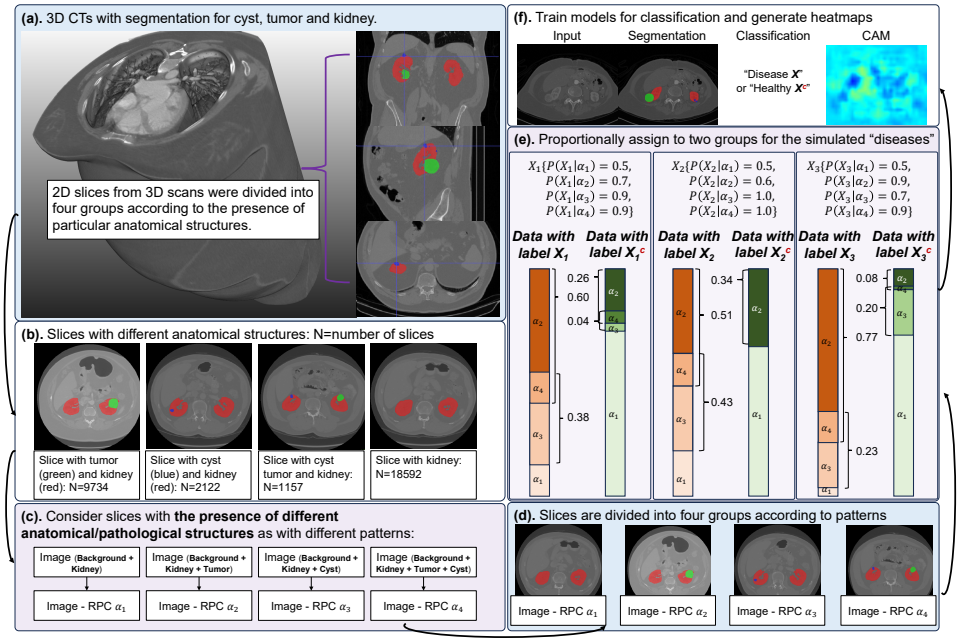


Figure 5.4: The process of generating virtual diseases $X_k, k \in \{1,2,3\}$ with different posterior probabilities of X_k given $\alpha_k, k \in \{1,2,3\}$. (a) CTs from KiTS23 with segmentation of kidneys, tumors and cysts. (b) Slices were divided into four groups according to the presence statuses of kidneys, cysts and tumors, N represents the number of slices containing particular anatomical/pathological structures. (c) The presence or absence of 'kidneys', 'tumors', 'cyst' and 'tumor and cyst' were considered RPC α_1, α_2, α_3 and α_4. (d) The slices are then divided into four groups according to patterns, and be assigned to the "disease X" and "healthy X^c". (e) The three simulated "diseases X" based on particular RPCs. On these three different simulated diseases, the trained models obtained AUROCs of 0.74, 0.72 and 0.78 close to their upper limits. (f) An example of the input, segmentation, classification and CAM for a task.

prior probability of X to be 0.5. For simple illustration and explanation, the $P(X|\alpha_2)$ and $P(X|\alpha_3)$ include the RPCs included in $P(X|\alpha_4)$, but are not what is calculated during the dataset generation. Three different diseases were designed to be:

- $X_1\{P(X_1|\alpha_1)=0.5, P(X_1|\alpha_2)=0.7, P(X_1|\alpha_3)=0.9\}$
- $X_2\{P(X_2|\alpha_1)=0.5, P(X_2|\alpha_2)=0.6, P(X_2|\alpha_3)=1.0\}$
- $X_3\{P(X_3|\alpha_1)=0.5, P(X_3|\alpha_2)=0.9, P(X_3|\alpha_3)=0.7\}$

RPC α_2 and α_3 have different marginal probabilities ($P(\alpha_k)$, $k \in \{2,3\}$) and RPC α_1 is set to be the reference as irrelevant RPC. Then three DL models with the same architecture were trained to discriminate $X_k, k \in \{1,2,3\}$ from healthy, and then

CAMs were generated and analyzed to obtain the mean activations in the regions of $\alpha_k, k \in \{1,2,3\}$ for validation. The whole process of generating virtual diseases $X_k, k \in \{1,2,3\}$ are presented in Fig. 5.4.

5.3.2 Simulation 2: quantitative evaluation

A series of virtual diseases $P(X_2^k|\alpha_2), k \in \{0.50 + 0.05n \mid n = 0,1,\ldots,10\}$, aiming at quantitatively evaluate the sensitivity and accuracy of the aggregation, were generated to validate if a change in the 'importance' -value is in accordance with a change posterior probabilities of X given RPC α_2. The posterior probabilities are $X_2^k\{P(X_1|\alpha_1) = 0.5, P(X_1|\alpha_2) = \boldsymbol{k}, P(X_1|\alpha_3$ where $k \in \{0.50 + 0.05n \mid n = 0,1,\ldots,10\}$.

5.3.3 Rheumatoid arthritis prediction

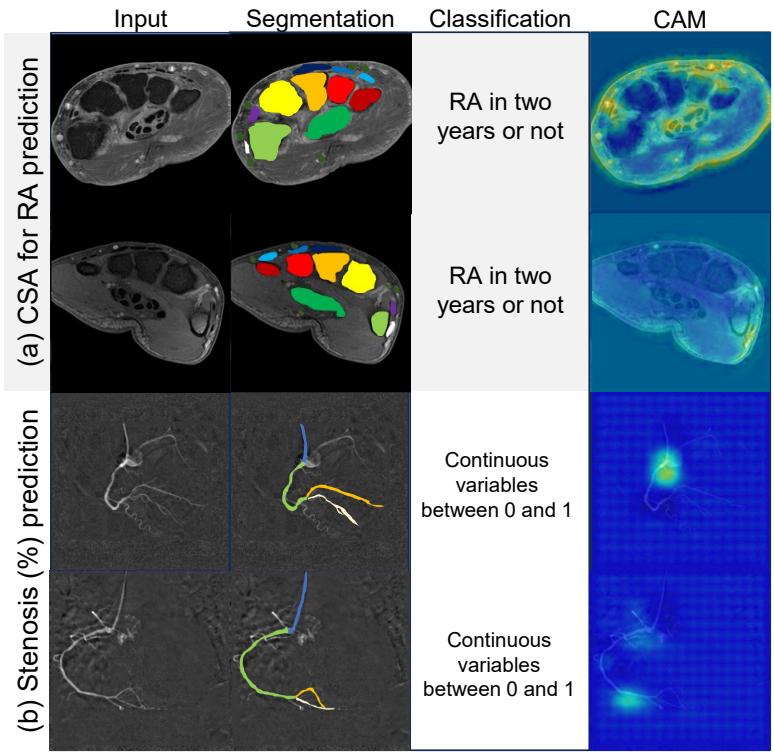


Figure 5.5: Illustrative examples of (a) MRIs from the CSA cohort for rheumatoid arthritis (RA) prediction, corresponding CAMs and segmentation generated by a trained segmentation model. (b) X-rays from Medis Medical Imaging Systems for stenosis percentage prediction, corresponding CAMs and manually created segmentation. Segmentation of these anatomical structures were considered as the image patterns that are potentially relevant to determine the outcomes for each task, and the locations containing these structures were considered the regions containing specific image patterns.

To check if the proposed aggregation method could perform and receive reasonable results on real clinical tasks, we applied the proposed aggregation method to a dataset for rheumatoid arthritis (RA) prediction. Ideally, the aggregated CAMs would assign high importance to the regions, of which the inflammation or other image features are predictive to RA.

In this study, wrist MRI scans of 727 patients with clinical suspect arthralgia [147], some of whom developed rheumatoid arthritis (RA) in two years, were used for this illustration. A deep learning model was previously trained based on this dataset to predict this development of RA [72] with an AUROC of over 0.7, which is close to the performance of the study based on visual scoring. CAMs with a global intensity scale [57] were generated to indicate the most informative regions to predict RA [72]. Meanwhile, the segmentation of wrist MRI scans from [11] employed as the region definition in this dataset. The segmentation consists of 33 different objects in the axial MRI scans of wrists, including 14 tendons, 15 bones, vessels, background, skin and remaining tissue. Based on these materials, the aggregation was set to assign an importance values to each anatomical regions that were labeled by the segmentation, and roughly validated by checking whether these importance values are reasonable according to the prior knowledge.

Fig. 5.5 (a) presents the examples of the input, segmentation, outputs and the corresponding CAM.

5.3.4 Stenosis score prediction task

Another dataset, for illustrating the potential use of the aggregation framework, was authorized by Medis Medical Imaging Systems. The aim of the DL models on this dataset was to estimate the stenosis scores (measurement for the percentage stenosis) based on X-ray angiograms for cardiovascular right coronary arteries (RCAs). We trained an end-to-end model to obtain stenosis scores percentage by inputting using the X-rays angiograms as input, with an AUROC of 0.8. In principle, the stenosis score percentage is defined as the diameter reduction ratio at the location with the most severe narrowing along the RCA in an X-ray angiogram. However, the accurate segmentations of narrowing areas are not available and therefore makes it difficult to check the alignment of the mean activations of narrowing regions and the areas of narrowing regions. Therefore, we used the manual segmentation of the vessels and check if the CAMs received higher signals in the regions of RCA instead of other regions. The segmentation consists of five classes, including RCA, RPD, RPL, catheter and background. Fig. 5.5 (b) presents the examples of the inputs, segmentation manually labelled by human experts, outputs and CAMs for illustration.

5.4 Experiments and results

5.4.1 Qualitative validation on the Simulation 1

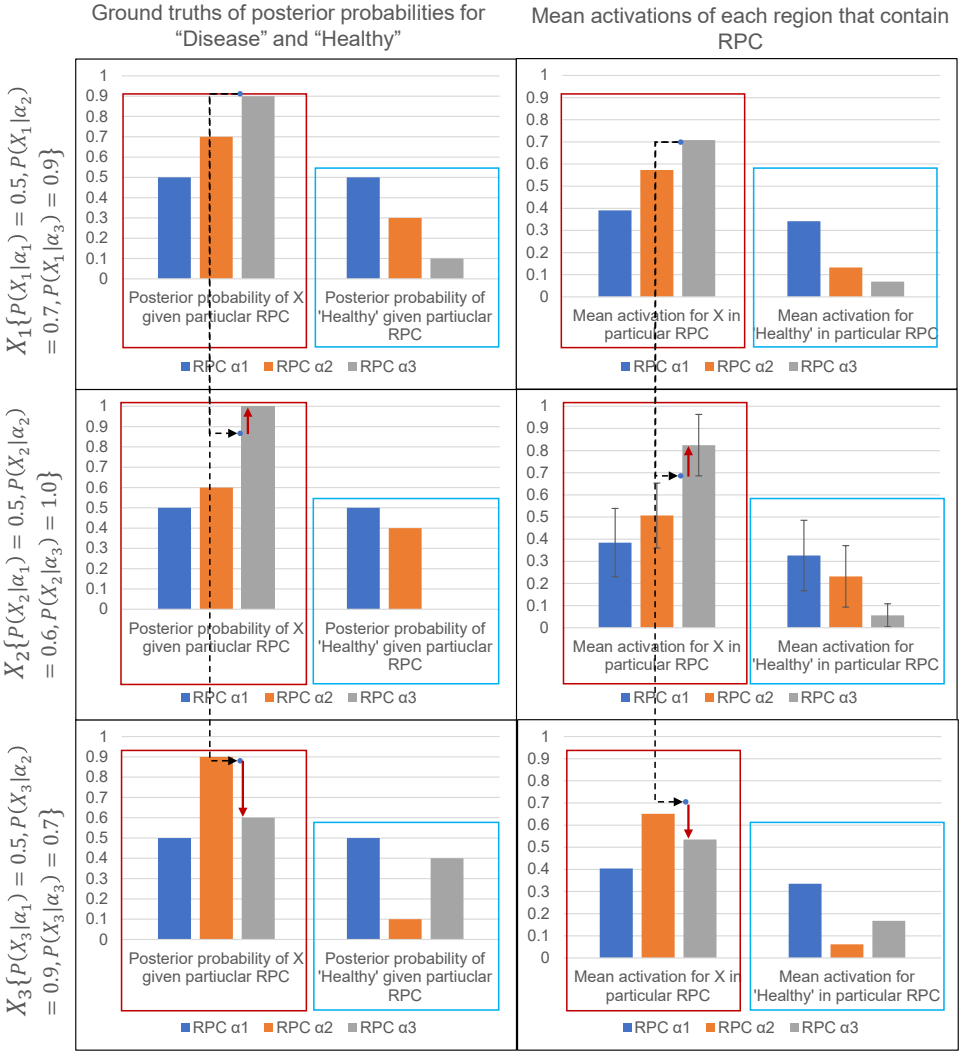


Figure 5.6: Results of the CAM aggregation of the classification tasks for diseases $X_k, k \in \{1,2,3\}$. Red boxes and blue boxes indicate the posterior probabilities and mean activations regarding disease "X" and "healthy" , respectively. Red arrows mark the changes in $P(X_k|\alpha_3)$ and mean activations in α_3 (compared to X_1), which showed high correlations with each other.

The models used for obtaining the CAMs were trained to have a performance very close to their upper limits (according to the posterior probabilities) to avoid potential errors

caused by poor-performance models. Likewise, the CAMs were generated through an optimal CAM algorithm from our previous study [57] that achieved the highest correlations between the models' outputs and the mean activations to minimize errors from the CAM algorithms.

Fig. 5.6 presents the results, comparing the importance-values from aggregation and the corresponding posterior probabilities. From X_1 to X_2, the posterior probabilities $P(X|\alpha_3)$ increased, $P(X|\alpha_2)$ decreased, $P(X^c|\alpha_2)$ decreased and $P(X^3|\alpha_3)$ increased. These changes in posterior probabilities were reflected in the changes in mean activations in the region of α_2 and α_3 according to the aggregation. Similarly, the difference between X_1 to X_3 were also found by the aggregation.

As the mean activations in the CAMs reflected the importance of these RPCs in the original images, this experiment qualitatively proves the reliability of the proposed CAM aggregation. Meanwhile, as can be seen from X_1 and X_3 that have an opposite posterior probability for α_2 and α_3, the disease "RPCs" are "equivalent" in the CAM aggregation – it equally reflects the contribution of these RPCs to the diagnosis without a significant preference for certain regions.

5.4.2 Quantitative validation on the Simulation 2

Tab. 5.2 presents how mean activations in all the regions when continuously change the ground truths for the posterior probabilities $P(X_2^k|\alpha_2), k \in \{0.50 + 0.05n \mid n = 0, 1, \ldots, 10\}$. While the mean activations of other regions in CAMs presents very few changes corresponding to the changes of posterior probabilities, the mean activations of α_2 in the CAMs increase and decrease accordingly to the α_2' s importance to X_2^k and to 'Healthy' with Rs of nearly 98% and -97%, respectively. This demonstrates that the changes in the mean activations in region α_2 reflect the changes of the posterior probabilities. Combining with the results of the qualitative evaluation, through the proposed CAM aggregation, the mean activations in CAMs is capable of revealing the 'importance' of an RPC to diagnose a disease.

Table 5.2: Mean activations for X and 'Healthy' with standard deviations

Posterior probability	Mean activations for X			Mean activations for Healthy		
of X_2 given α_2	$\alpha_1 \pm$ std	$\alpha_2 \pm$ std	$\alpha_3 \pm$ std	$\alpha_1 \pm$ std	$\alpha_2 \pm$ std	$\alpha_3 \pm$ std
0.5	0.2650 ± 0.0313	0.2624 ± 0.0338	0.2695 ± 0.0376	0.1998 ± 0.0540	0.2047 ± 0.0652	0.2029 ± 0.0663
0.55	0.3487 ± 0.0689	0.3738 ± 0.0828	0.3223 ± 0.0815	0.2435 ± 0.0609	0.2317 ± 0.0676	0.2530 ± 0.0806
0.6	0.3613 ± 0.0555	0.4094 ± 0.0667	0.3358 ± 0.0684	0.2685 ± 0.0494	0.2232 ± 0.0582	0.2991 ± 0.0618
0.65	0.3024 ± 0.0347	0.3953 ± 0.0322	0.2472 ± 0.0223	0.2935 ± 0.0156	0.1894 ± 0.0220	0.3134 ± 0.0168
0.7	0.3894 ± 0.0798	0.5098 ± 0.0990	0.3481 ± 0.1014	0.2342 ± 0.0823	0.1513 ± 0.1048	0.2727 ± 0.0904
0.75	0.3785 ± 0.0432	0.5100 ± 0.0522	0.3383 ± 0.0579	0.3255 ± 0.0620	0.1187 ± 0.0695	0.3544 ± 0.0767
0.8	0.3480 ± 0.0606	0.5655 ± 0.0732	0.3142 ± 0.0794	0.2923 ± 0.0349	0.1278 ± 0.0422	0.3351 ± 0.0422
0.85	0.3562 ± 0.0520	0.5652 ± 0.0654	0.3255 ± 0.0713	0.2856 ± 0.0372	0.0901 ± 0.0407	0.3134 ± 0.0403
0.9	0.3833 ± 0.0508	0.6606 ± 0.0620	0.2857 ± 0.0682	0.3293 ± 0.0480	0.0627 ± 0.0662	0.3821 ± 0.0567
0.95	0.3465 ± 0.0502	0.6250 ± 0.0597	0.3151 ± 0.0653	0.2457 ± 0.0622	0.0254 ± 0.1022	0.2564 ± 0.0756
1	0.3534 ± 0.0565	0.7225 ± 0.0660	0.3416 ± 0.0722	0.2866 ± 0.0497	0.0122 ± 0.0795	0.2759 ± 0.0632
R	0.4701	0.9736	0.2837	0.5102	-0.9726	0.4080

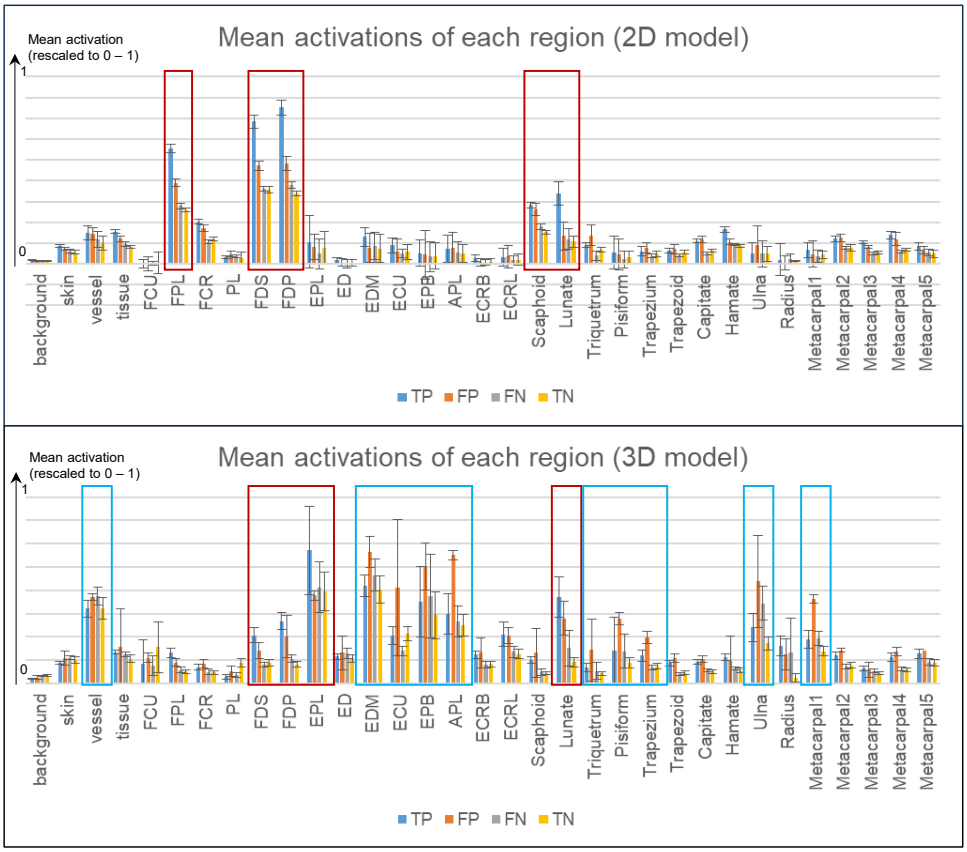


Figure 5.7: Results of the CAM aggregation on RA prediction task, based on MRIs. (a) The results based on 2D models and (b) the results based on 3D models. The red boxes highlight regions with higher mean activations for true positive cases, while the blue boxes highlight those with higher mean activations for false positive cases (misleading regions). The disagreement between these two models on the importance of regions (containing image patterns), which may be caused by segmentation and CAM(2D)-segmentation(2D/3D) aggregation process (see the section on method), nonetheless indicates the potential bias of CAM aggregation due to different models.

5.4.3 Application to rheumatoid arthritis prediction task

We applied the CAM aggregation to the model for predicting RA based on MRIs, where this aggregation could generate some fully data-driven hypotheses on which features are more informative among all features, which induces further explorative studies in the future. To minimize potential biases caused by the model, five-fold cross-validation was applied and the mean results on a hold-out test set were presented. Meanwhile,

another model trained through a different model architecture, dataset split, and input dimension was applied as a comparison. Fig. 5.7 presents the mean activations in CAMs for each region according to deep learning-based segmentation on the MRIs of wrists.

According to the trained models with current performance and the dataset, despite the disagreement on the importance of vessels, FPL, EPL and scaphoid, regions around FDS, FDP and Lunate obtained higher mean activations compared to other tendons, bones or tissues. Furthermore, the results can be very different, as can be seen from the difference between the two models, based on the same dataset, different data splits and aggregation methods (due to 2D or 3D segmentation). Therefore, potential biases or errors should be noticed while analyzing the CAM aggregation results. In summary, this generates a fully data-driven hypothesis, which needs further investigations and studies from a clinical perspective.

5.4.4 Application to predicting stenosis score

Tab. 5.3 presents the mean activations for predicting stenosis score from the X-rays, in which segmentations of background, catheter, RCA, RPD and RPL were provided. According to the definition of the stenosis scores of this dataset, regions of RCA that may contain patterns of stenosis should obtain the highest mean activations for predicting stenosis scores. However, the CAM aggregation based on the trained model indicates that RPD and RPL are as informative as RCA.

As mentioned in the section on datasets, the relationship between the stenosis scores and the narrowing regions in RCA is clear. Therefore, if we aimed to mimic the standard calculation process of stenosis scores, the model failed to match the requirement of measuring stenosis scores based on RCA, namely this is not a reliable model. From another perspective, this result suggests that RPL and RPD are also informative in measuring stenosis scores, providing a hypothesis that the stenosis in RCA may affect downstream vessels of RPD and RPL, making it possible to estimate the effect of stenosis in these structures.

Table 5.3: CAM aggregation results on stenosis score prediction, using a trained model and a randomly initialized model. Compared to randomly initialized model, in the trained model, the RCA, RPL and RPD received relatively high values in CAMs compared to background and catheters.

Regions	Trained model	Randomly initialized model
Background	0.18(±0.17)	0.28(±0.21)
Catheter	0.19(±0.16)	0.36(±0.17)
Right coronary artery (RCA)	0.34(±0.11)	0.35(±0.24)
Right posterior lateral branch (RPL)	0.33(±0.12)	0.31(±0.27)
Right posterior descending artery (RPD)	0.34(±0.14)	0.29(±0.17)

5.5 Discussion

The proposed CAM aggregation provides a method to (1) automatically validate whether models are in alignment with expert knowledge, if the relationship between regions containing specific image patterns and the models' outputs is clear; (2) generate hypotheses about which regions contribute more to the outputs, revealing the correlation between new image patterns and clinical outcomes. However, some factors have an impact on the accuracy of the aggregation – the bias of models/datasets and the accuracy of segmentation.

5.5.1 Bias of models and datasets

The proposed CAM aggregation is a method that focuses mostly on interpreting the deep learning models, instead of the datasets. Therefore, when the model goes wrong, the conclusion that a certain region with image patterns is informative may also be misleading (See the results of using different models for RA prediction). Models with a high performance can help mitigate misleading information caused by some bias of the models, yet high-performance models are not always available in medical tasks. A possible solution is using models trained with different methods to achieve a cross-model CAM aggregation thereby minimizing the bias of a single model. However, the bias of datasets is almost unavoidable, especially for small datasets, like in the medical field. Therefore, the conclusion of CAM aggregation on a single dataset is a conclusion on this specific dataset and is just a hypothesis for other datasets of the same topic, just as the situation for any DL models and methods developed for a specific dataset. Therefore, in general, CAM aggregation serves more as a hypothesis generator and provides possible candidates of regions of patterns for further investigation.

5.5.2 Accuracy of segmentation

As we take semantic segmentation to find the locations of patterns in the original images, the accuracy of the segmentation labels is crucial for a correct conclusion. Errors in the segmentation can lead to over- and underestimation of certain regions containing image patterns, especially when there are overlaps of two segmentation labels (two patterns) in a region. The aggregation results of using different models for RA prediction may differ from each other due to the biases of models or the errors in segmentation.

5.5.3 Definition of importance

In this study, the mean activation in a region was considered the importance of that region. However, the spread of local activation (by reporting maximum, minimum and standard deviation) can also provide additional insight into the process of a DL model and the distribution of the features in the dataset. For example, if an inflammatory sign rarely appears in a particular anatomical region across all MRI scans, but this

inflammation is very prediction. The mean activation of this inflammation may be low and not considered as important image pattern according to current definition. In this case, current definition of 'importance' using mean activations is not a comprehensive measurement of the real importance. The solution to this problem is to have more than just mean activations included, but also other information like the spread, standard deviations.

In summary, current definition of the importance for each region is merely a measurement that emphasizes regions frequently highlighted in CAMs. According to the purpose of using the proposed CAM aggregation, the importance can be defined in different ways.

5.5.4 CAM aggregation: a qualitative evaluation

An essential remark while interpreting the outputs of CAM aggregation is that **the absolute mean activations are highly sensitive to the CAM algorithms, models and CAM intensity scaling, and therefore comparison of the absolute values between different models are meaningless**. However, for the combination of a specific model, CAM algorithm and CAM aggregation, the mean activations among different regions are comparable under the same intensity scaling. Therefore, the relative relationship among the mean activations of these regions is the most important. That means, CAM aggregation provides a qualitative evaluation rather than quantitative of which regions with what patterns are more informative to the output.

5.6 Conclusion

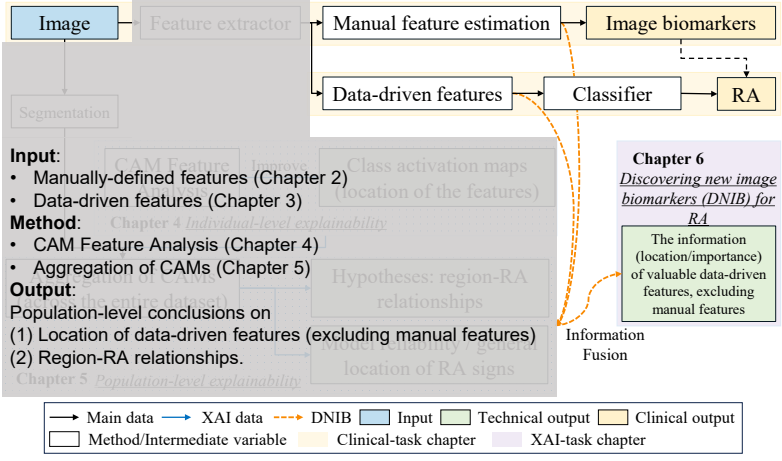
In this study, we proposed a method to aggregate the information in CAMs to achieve a population-level analysis. This method can serve as a validation method for deep learning models provided that the relationship between regions that contain specific patterns and outcomes is clear. Otherwise, CAM aggregation provides a fully data-driven method of comparing the importance among regions of patterns and revealing unknown correlations between regions and clinical outcomes.

Acknowledgment

This work is supported by the Netherlands Organization for Scientific Research (NWO, TTW 13329), the European Union's Horizon 2020 research and innovation programme (No.714312) and the China Scholarship Council (No.202108510012). The datasets are supported by Leiden University Medical Center and Medis Medical Imaging Systems.

6

Auxiliary-branch CAM in deep learning models serves as a tool for discovering undefined image patterns of rheumatoid arthritis



This chapter was adapted from:

Yanli Li, Denis P. Shamonin, Monique Reijnierse, Annette H.M. van der Helm-van Mil and Berend C. Stoel. "Auxiliary-branch CAM in deep learning models serves as a tool for discovering undefined image patterns of rheumatoid arthritis." (*in preparation*)

Abstract

Deep learning (DL) methods have been rapidly developing in the field of medical image analysis. Based on different types of inputs and outputs, DL methods are typically applied to documentation, treatment planning, image-guided intervention and diagnosis. In this study, we explored a different application of DL – discover undefined image patterns that could contribute to rheumatoid arthritis (RA) prediction, based on a multi-task model architecture, a revised class activation mapping (CAM) algorithm, the segmentation-based aggregation of CAM and information flow control. We propose a CAM-based image pattern discovery (CAMPID) framework, which split the forward of a DL model into a primary path for estimating manually-defined image patterns and an auxiliary branch for extracting undefined image patterns. Through generating and aggregating the auxiliary-branch CAMs, this framework assigns a 'importance' value to each region that represents its contribution to RA prediction. By comparing the 'importance' value in each region with background and irrelevant objects, the undefined yet predictive image pattern can be then located. We validate this framework using a dataset that has records of RA development and manually-defined inflammatory sign scores, moving one of the manually-defined inflammatory signs in and out of the primary path. The experiment results demonstrate the feasibility of this framework, as the 'importance' value in the regions of the inflammatory signs changed accordingly to their information existence in the auxiliary branch. We expect this study could provide an initial idea for a novel application of DL models in medical image analysis.

6.1 Introduction

With the rapid development of artificial intelligence, particularly deep learning (DL) techniques, the reliable and explainable application of DL models has the potential to significantly accelerate and enhance medical image processing and analysis. Leveraging large-scale data, DL models with strong cross-center robustness and interpretability have become a key focus in the field of computer-assisted intervention. These models are expected to offer more intelligent, efficient tools for analysis, clinical assistance, and diagnosis in future medicine.

In this context, current DL models are typically applied in areas such as clinical communication and documentation (e.g., large language models, LLM) [148, 149], treatment planning and image-guided intervention (e.g., segmentation) [150, 151], survival prediction [152] and diagnostic classification tasks [153, 154]. These applications aim to model and learn the relationships between inputs and outputs through data-driven, automated approaches. Many DL methods are designed to emulate expert reasoning and decision-making processes, as this allows for step-by-step validation of the model's alignment with expert knowledge, thereby facilitating the development of more reliable DL models. For example, DL models in medical imaging are often developed to estimate intermediate biomarkers or verifiable features, such as cancer-related scores or biomarkers [155, 156], tumor voxels [157], segmentation for radiation therapy planning [158].

At the same time, many DL studies have aimed to take advantage of the data-driven nature in DL to discover informative features independent of clinical hypotheses. In this regard, end-to-end DL models, which directly map raw medical images to clinical outcomes, have shown remarkable success in diagnostic applications [159, 160, 161]. These achievements demonstrate the capacity of DL models to extract clinically relevant information from inputs without relying on expert-defined features. Furthermore, the implicit representations learned by these models may offer new perspectives for validating and re-evaluating human-defined imaging patterns (i.e., biomarkers), thereby supporting both technical innovations and clinical research.

To interpret the implicit representations learned by DL models, interpretation methods like class activation mapping (CAM) are proposed to locate the informative image patterns and compare them with expert knowledge. Through CAM algorithms, attention maps (i.e. heatmaps, saliency maps) are generated, highlighting regions that substantially contribute to DL model's diagnoses. By visually checking these highlighted regions, researchers can assess the consistency between the DL model's focus and established expert knowledge. Such CAM-based interpretability techniques have become common practice in recent medical imaging studies [162, 163, 164].

During such visual check, if the highlighted regions in CAMs contradict expert-

defined relevant areas, the model is often suspected to be influenced by confounding variables or misleading factors. However, when the presence of such confounders can be reasonably excluded, these seemingly 'incorrect' regions may actually reveal previously undefined image patterns associated with the clinical outcomes. Inspired by this concept, we investigate a potential method – using DL to discover undefined image features that contribute to the prediction of rheumatoid arthritis (RA).

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disorder that primarily affects the joints in the wrists, hands, and feet [3]. Inflammatory signs visualized through magnetic resonance imaging (MRI) have been shown to be predictive of RA development. Among these signs, several imaging biomarkers, specifically bone marrow edema (BME), synovitis (SYN), and tenosynovitis (TSY), have been defined by clinicians as important indicators for RA prediction [80, 82].In our previous work, we developed an end-to-end deep learning (DL) model to predict RA development directly from MRI scans, demonstrating the model's ability to capture RA-related imaging patterns. In a separate study using the same dataset, we further showed that DL models can estimate the presence and severity of inflammatory signs at a level comparable to that of human experts, effectively extracting information from MRI scans aligned with clinical definitions.

With the help of class activation mapping (CAM) algorithms and anatomical knowledge, human experts can visually inspect whether the RA prediction model bases its decisions on these predefined inflammatory signs. If the average activations within regions corresponding to these signs are significantly higher than those in irrelevant areas, it supports the model's ability to autonomously discover the relationship between these features and RA, even without explicit expert guidance. Similarly, other previously undefined image patterns related to RA might be uncovered if they also exhibit high CAM activations. However, while this idea of identifying novel image patterns is intuitive and straightforward, several factors – such as L1 and L2 regularization, which constrain models to focus on the most salient features – may cause the end-to-end DL model to overlook less prominent but potentially informative patterns.

To address this limitation, we propose a multi-task learning framework, CAM-based image pattern discovery (CAMIPD), which integrates the end-to-end RA prediction model with the inflammation (inflammatory sign) estimation model. By modifying the forward and backward propagation mechanisms, we regulate the flow of information within the model. Specifically, this framework establishes two distinct forward paths:

- A primary path that estimates the predefined inflammatory signs, which are then used as input to a classifier for RA prediction. Importantly, the backward path from RA prediction to inflammation estimation is detached to prevent gradient

flow.

- An auxiliary path that captures additional, potentially predictive information beyond the predefined signs. Here, the backward path is retained, allowing the model to optimize for undefined image patterns relevant to RA prediction.

Using the same dataset as in our previous studies, we validate this framework by examining whether image patterns not included in the main estimation path can still achieve high mean activations in the CAMs derived from the auxiliary path.

The layout of this paper is as follows. First, we introduce our MRI materials and define the task. Subsequently, we clarify the proposed framework and the corresponding architecture of the DL model, accompanied by the method for validation. We then present experiments evaluating the proposed method. Finally, we discuss and summarize the limitations and advantages of the proposed framework.

6.2 Material and method

6.3 Material

To develop and validate the proposed method for discover undefined image patterns, we utilized a database [147, 72] that has labels of RA development and scores of inflammatory signs in different regions. In total, wrist MRI scans from 616 CSA patients with mean age (first time scanned) of 43.2 ± 12.6 years, 76.8% female ratio, 2 scores (spread: 1 to 5) of joint inflammation under visual scoring system [22] were adopted for model training and CAMIPD. Informed consent was given by all patients, referring to LUMC protocol reference numbers: B19.008 and P11.210.

A visual scoring system based on the RAMRIS criteria [22] was used to evaluate each anatomical region. Two trained readers, blinded to clinical data, independently scored the images. For tenosynovitis (TSY) and synovitis (SYN), scores ranged from 0 to 3, reflecting the estimated maximum width of peri-tendinous and synovial effusion/proliferation (respectively), observed with contrast enhancement [22]. The grading scale was defined as follows: grade 0 to normal; grade 1 to $\leq$ 2 mm; grade 2 to $>$ 2 mm and $\leq$ 5 mm; and grade 3 to $>$ 5 mm. The scoring area extended from the distal radius/ulna proximally to the hook of the hamate distally. Bone marrow edema (BME) was also assessed on a 0 to 3 scale, based on the estimated proportion of the bone volume affected: grade 0 for no edema; grade 1 from 1 to 33%; grade 2 from 34 to 66%; and grade 3 from 67 to 100%. The average of the two readers' scores was considered as the ground truth reference for model training and CAMIPD.

The patients in this dataset were followed over a period of two years in order to establish whether they had developed RA, and the status is then considered as the label of RA development in this study. The technical information can be seen in 6.1.

Table 6.1: Technical parameters of the hand and forefoot MRI scans.

MRI Parameter	Transversal (axial) scan	Coronal scan
In-plane matrix	320x192	364x224
Repetition time	570 msec	650 msec
Echo time	7 msec	17 msec
Echo train length	2	2
Slice thickness	3 mm	2 mm
Slice gap	0.3 mm	0.2 mm
Fat saturation	Frequency-selective fat saturation applied	
Scanner	1.5T extremity MRI scanner (GE Healthcare) using a 100-mm coil	

The 616 MRI scans were split into five folds for five-fold cross-validation. In each fold, three folds were used as training set, one fold was retained as test set for model training and one fold served as the validation set.

6.3.1 General workflow

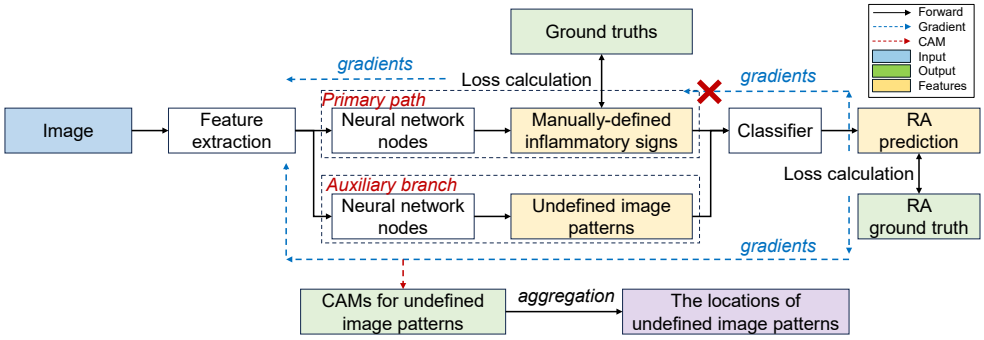


Figure 6.1: The general workflow of the proposed CAMIPD framework for discovering undefined image patterns. The forward path of the DL model is split into a primary path for learning manually-defined inflammatory signs and an auxiliary branch to capture undefined image patterns that are predictive to RA prediction.

Fig. 6.1 presents a general workflow of the proposed CAM-based image pattern discovery framework. To obtain 'pure' CAMs only highlight undefined image patterns, the forward path of the DL model is split into two branches: (1) a primary path for learning the manually-defined inflammatory signs and minimizing the information of these predefined features in other branches; (2) an auxiliary branch to extract the undefined image patterns and generate CAMs based on the gradient of this branch. The gradient backward calculation from the primary path to classifier is detached to constrain primary path to focus on manually-defined inflammatory signs and avoid information leakage – undefined image patterns may also go through the primary path under the guidance of gradients.

6.3.2 Model architecture

Table 6.2: Configuration of model architecture and training

Modules & Hyperparameters	Configs
3D Conv Block (1st block)	Conv: depth: 2, kernel size: [3×3×1], number of kernels: 32, stride: 1, group: 1/2 (view), Activations: ReLU.
3D Conv Block (2nd – 4th block)	Conv: depth: 2, kernel size: [3×3×1], number of kernels: 64, stride: 1. Pool: pooling size: [3×3×1], max pool, stride: 2, group: 1/2. Activations: ReLU.
3D Conv Block (5th – 6th block)	Conv: depth: 2, kernel size: [33×1], number of kernels: 96/160 (5/6th), stride: 1, group: 1/2. Pool: pooling size: [3×3×1], max pool, stride: 2. Activations: ReLU.
a Cross-attention block (7th block)	In and out channel: 160. Inner dense layer channel: 640. Window size: [2×2], Head: 4, Dimension of head: 8. Dropout rate: 0.0.
Adaptive average pooling	Output size: [1×1×1].
Dense layer	Layer 1: [640×1/2 (views)×1×1×1, #scores + #additional]. Layer 2-1: [#scores, #hidden nodes]. Layer 2-2: [#additional, #hidden nodes]. Layer 3: [#hidden nodes, 2 (RA labels)].
Optimizer	AdamW.
Learning rate	4e-5.
Weight decay	1e-2.
Optimizer momentum	0.9, 0.999.
Training epochs	120.
Gradient clip	None.

The DL model in [72] is adopted for this study, with a series of convolutional layers and transformer blocks as feature extraction, a layer of nodes to contain the values of learned features and a two-layer dense layer as classifier. The manually-defined inflammatory signs and undefined image patterns are split at second layer of dense layers, go through different paths and are then merged as the inputs of the classifier. The details of the model configuration and hyperparameters are shown in Tab. 6.2. The model architecture can take two different input – one-view transversal (TRA) MRI scans or two-view TRA and coronal (COR) MRI scans. Both input modes were experimented in this study to validate the robustness. This model structure and training protocol have achieved performance close to the level of human experts combined with statistical models in RA prediction and inflammation estimation according previous studies.

6.3.3 Class activation mapping and aggregation

A revised version [57] of the CAM technique [91, 58, 55] is applied to generate the CAMs for the undefined image patterns. This revised CAM method allows for fair comparison within a dataset, by building a global intensity scale for normalization. Subsequently, we use segmentation [11] to define the regions in the original images and aggregate the CAM activations in each region. This segmentation-based aggregation will assign an 'importance' value to each region, which represents its contribution to model' s outputs. The higher is the 'importance' value, the more predictive is the image pattern in the region.

6.3.4 Validation method

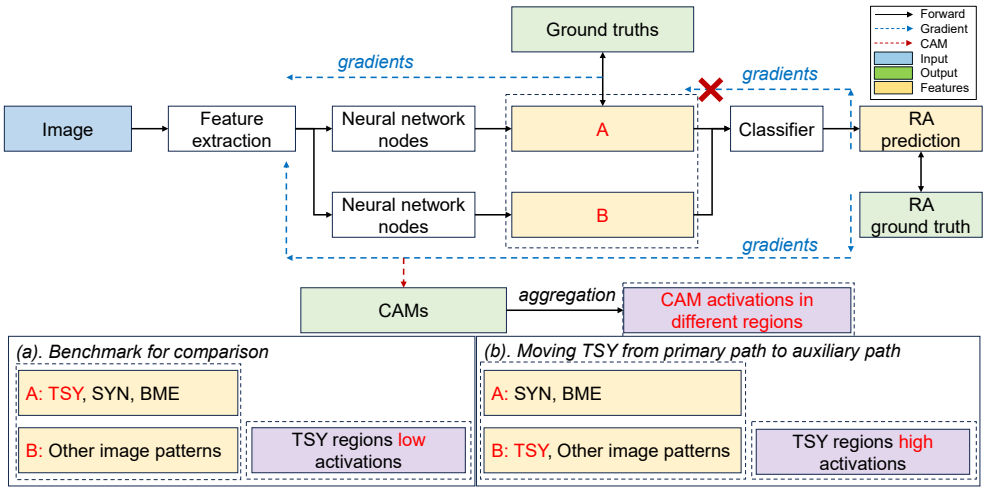


Figure 6.2: Validation method through comparing CAMs based on comparing the CAM activations in TSY regions before (a) and after (b) moving TSY from primary path to auxiliary path. By filling A, B with different setting in (a) and (b), the CAM activations in TSY regions could vary and have different activation level compared to other regions. Since the auxiliary-branch CAM is designed to visualize the information flow through auxiliary branch, the CAM activations in TSY regions are expected to increase substantially after moving TSY from the primary path to auxiliary path.

If the proposed CAMIPD framework is a valid method, the regions of predictive image patterns should have higher CAM activations compared to the regions of background or irrelevant objects. This validation requires at least one predictive image patterns that are not included in manually-defined inflammatory signs, which is not feasible according to current studies on RA prediction. Therefore, we changed the thoughts and put a manually-defined inflammatory sign, TSY which is most predictive to RA in this specific study, out of the primary path. After this removal, the CAMs for the

auxiliary branch should highlight the regions of TSY more than before this removal, as TSY has become the 'undefined' predictive image pattern (see Fig. 6.2). The process is applied to the two models with input of one- (only TRA scans) and two-view (TRA and COR scans) MRI scans. The mean activations in all regions in segmentation will be presented and compared to check whether the mean activations of TSY regions increased after removal from the primary path.

6.4 Experiments

6.4.1 Backbone model performance

The multi-task models, using one-view (TRA) and two-view (TRA and COR) MRIs as inputs, received a comparable but a little lower performance on both RA prediction task and inflammation estimation task compared to the original performance before applying multi-tasking (see Tab. 6.3). The decrease of RA prediction accuracy derived from multiple factors: (1) dropout function was removed from the model to purify the information in each node, but dropout could restrict overfitting and improve model's performance and robustness. (2) Some of the manually-defined inflammatory signs may affects the prediction and cannot be neglected as the backward path is detached. (3) The number of 'free' hidden nodes that can be trained using gradients from RA prediction decreased as some nodes were used for inflammation estimation, leading to less trainable parameters for RA prediction. The decrease of inflammation estimation correlation may originate from similar reasons. Nonetheless, the DL models in practice are rarely perfect, the little decreases in performance ought not to affect a robust method.

Table 6.3: The performance of the multi-task models used for the CAM-based image pattern discovery and the original performance of the same model architecture but not multi-tasking in previous studies (the benchmark).

Model performance on each task after multi-task training						
Setting	**Input**	**TSY in primary path?**	**RA AUC**	**RA F1**	**Inflammation R**	**Inflammation ICC**
Multi-task (this study)	TRA scans	True	0.68 (± 0.01)	0.68 (± 0.02)	0.83 (± 0.01)	0.79 (± 0.02)
	TRA scans	False	0.67 (± 0.02)	0.67 (± 0.03)	0.77 (± 0.01)	0.68 (± 0.04)
Multi-task (this study)	TRA and COR scans	True	0.71 (± 0.02)	0.71 (± 0.01)	0.81 (± 0.01)	0.66 (± 0.03)
	TRA and COR scans	False	0.72 (± 0.01)	0.73 (± 0.01)	0.82 (± 0.02)	0.67 (± 0.03)
Backbone model performance on each task in previous studies						
RA prediction	TRA scans	N/A	0.69 (± 0.05)	0.72 (± 0.04)	N/A	N/A
RA prediction	TRA and COR scans	N/A	0.73 (± 0.04)	0.72 (± 0.06)	N/A	N/A
Inflammation estimation	TRA scans	True	N/A	N/A	0.82 (± 0.02)	0.78 (± 0.02)
Inflammation estimation	TRA scans	False	N/A	N/A	0.78 (± 0.03)	0.67 (± 0.07)
Inflammation estimation	TRA and COR scans	True	N/A	N/A	0.86 (± 0.03)	0.75 (± 0.01)
Inflammation estimation	TRA and COR scans	False	N/A	N/A	0.81 (± 0.02)	0.74 (± 0.04)

6.4.2 Validation of the framework

Fig. 6.3 presents the mean activations in the segmentation-available regions of wrist MRI scans, based on one- or two-view models and TSY included in the primary path or not. As discussed in the method section, if TSY is not included in the primary path,

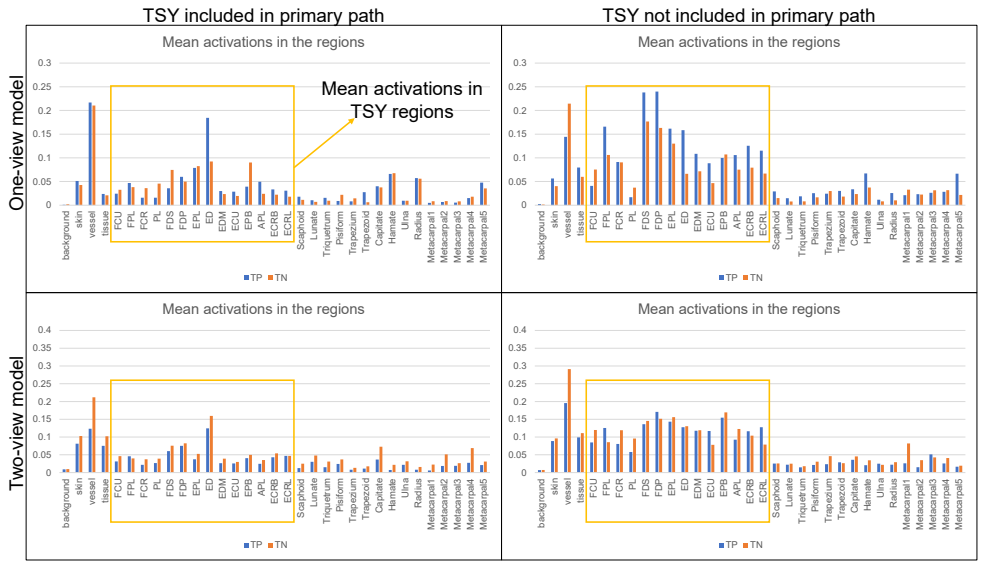


Figure 6.3: Mean activations (importance) in the regions of wrist MRI scans, based on one-view or two-view models and TSY included in primary path or not. Generally, in both one-view and two-view models, the mean CAM activations in TSY regions are higher when the TSY scores are not included, while other regions remained at a similar activation level.

the information of this predictive inflammatory sign must go through the auxiliary branch, otherwise the performance of the model would decrease due to the loss of this important information. As the consequence, when the model performance was not decreasing, their mean CAM activations in the auxiliary-branch CAM are expected to increase as their contribution to RA prediction through this path increased. As shown in Fig. 6.3, the TSY regions received more mean activations when the TSY is not included in the primary path, which match the expectation. These results demonstrate the feasibility of the proposed CAMIPD framework.

However, some the TSY regions received high activations even when TSY is included in the primary path. Several factors could lead to this phenomenon. (1) The most probable reason is that the inflammation scores based on the manually-defined scoring system are not the only predictive image pattern in the TSY regions – some additional signs could help to predict RA. (2) At the same time, the semi-quantitative feature of the manually-defined scoring system could also lead to incompleteness of describing TSY based on the system, resulting in residual TSY scores been highlighted in the auxiliary branch. (3) The inflammation estimation model is not perfectly aligned with the ground truths of inflammation scores. (4) The segmentation of some tendons, based on a previous study [11] may not be accurate and caused propagation of the

errors.

Meanwhile, as can be seen, the regions of vessels and skin continuously obtained high mean activations compared to other anatomical regions. This indicate that some signals or image pattern within these regions contributed to RA prediction and the information flows through the auxiliary branch. This could be some real image patterns or just confounding factors. In the wrist MRI scans, some high-intensity artifacts frequently occurred around the skin and the signal intensities of vessels are typically high. These high-intensity signals can be misinterpreted by DL models as inflammation, since inflammatory regions are also of high intensities.

6.5 Discussion

The experiments and results showed that the fundamental idea of the proposed CAMIPD framework is feasible. However, many limitations are substantial, limiting the use of the method and requiring more future investigations. The major factors that affect the accuracy of the proposed method are: (1) The accuracy of CAM algorithms; (2) The accuracy of segmentations; (3) The accuracy of the multitask model; (4) The prevalence and specificity of the undefined image patterns; (5) The purity of the information flow in auxiliary.

The accuracy of CAM algorithms, which is the source measurement for the 'importance' of an image pattern to RA prediction, could substantially influence the accuracy of the proposed CAMIPD framework. Since the proposed CAMIPD framework is calculated across different inputs, recording and comparing among the CAMs of these inputs, the CAM algorithm for this framework is supposed to have a global intensity scale and retain all available activations to avoid information loss. In this context, the method for building a global intensity scale may visually affect the numerical range of the mean activations in anatomical regions, leading to omission and even misinterpretation of some image patterns.

Similarly, the accuracy of segmentation affects the CAMIPD framework as it relies on segmentation to project CAM activations to the corresponding anatomical structures. A potential reason, for which some TSY regions obtained high CAM activations when their information ought to flow through the primary path, is the segmentation of these regions may mistakenly expand to other regions (e.g., regions of vessels). The solution to this problem is to have a better segmentation or manually check whether the high activations in some regions derive from wrong segmentation.

The accuracy of multitask model is similar. If the subtask of estimating image pattern scores failed to achieve a satisfactory performance, the primary path may not effectively extract the manually-defined image patterns, resulting in the auxiliary branch highlighting these regions as well. If the final outputs (e.g., RA prediction in this study) are correctly classified, the regions highlighted in CAMs may not point to

the expected image patterns, either.

Besides the potential errors caused by inaccurate upstream methods, the most influential factors are the prevalence and specificity of the undefined image patterns. The undefined image patterns must be prevalent in the entire dataset and informative enough to the final outputs (e.g., RA prediction in this study), so that they can be learned by DL models and captured by the mean activations of CAMs. If the image patterns are not prevalent, even if they are informative and receive highest CAM activations in some cases, they could be neglected as the mean activations could be low. At the same time, if the image patterns are not informative enough, the mean activations of these image patterns not stand out compared to other irrelevant objects or confounding factors.

A major limitation and issue to be solved in this CAMIPD framework is the information purity of the auxiliary branch. We chose to abandon dropout in the DL model, as dropout randomly masks part of the nodes to overcome the fitting and therefore could force some information in the primary path to flow through the auxiliary path. Consequently, the auxiliary-branch CAMs may highlight these regions even if there are no remaining undefined image features. Dropout is not the only factor that could 'pollute' the auxiliary branch, the relationship between the two tasks in the multitask model is another example. If the defined image patterns estimated through the primary path are not the only image patterns in the anatomical structures that contribute to the final output (e.g., RA), the mean activations in these regions could be high due to the undefined image patterns. In this case, it is difficult to determine the cause of the high mean activations in these regions – it can be meaningful image patterns, confounders, failure of information decoupling based on current model architecture or segmentation errors.

In summary, the CAMIPD framework with so many limitations unsolved in this study is merely an attempt to prove the feasibility of using DL models to discover undefined image patterns that could help predict RA. On the one hand, the validation qualitatively proves the feasibility of the framework to some extent. On the other hand, more quantitative evaluation based on more comprehensive and systematic experiments must be done in the future, alongside with the optimization of segmentation, model architecture and other factors.

6.6 Conclusion

In this study, we proposed a CAM-based image pattern discovery framework, aiming at a potential data-driven method to find the locations of particular image patterns that contribute to RA prediction. The experiment indicates the feasibility of this framework, with major limitations to be overcome and better evaluation method to be designed. We expect this study could provide an initial idea for a novel application of DL models

in medical image analysis.

6.7 Acknowledgments

This author was supported by the China Scholarship Council No.202007720110 during the development of this package.

7

Summary, discussion and future work

In this study, we have developed a series of models and methods for quantifying inflammation, predicting future RA development, interpreting DL models, aggregating information to reach population-level conclusions, and combining all these methods to discover potential new imaging biomarkers that contribute to RA prediction. In this chapter, we briefly summarize the previous chapters and discuss their advantages, limitations and potential future research directions.

7.1 Summary

In this thesis, the first chapter provides a general introduction to the research topic, the motivation and challenges to be overcome, and the relationship among the consecutive chapters. The initial idea of the whole research was to develop a data-driven method, independent of expert knowledge, to detect or predict RA at an early stage and also to find early signs of RA. The whole idea is divided into four objectives/steps: (1) Chapter 2: A potential application of DL models in clinical practice, and the feasibility of DL models in recognizing manually defined image biomarkers - this enables the potential discovery of additional image biomarkers other than these predefined image biomarkers; (2) Chapter 3: the feasibility of a method for predicting future RA development in MRI using DL models; (3) Chapters 4 and 5: a general framework to prove that the DL models are reliable and explainable, and whether they are reasoning based on some additional image biomarkers other than human experts, at individual and population level; (4) Chapter 6: A feasible framework for discovering new image biomarkers based on CAM algorithms.

In **Chapter 2** we investigated a practical application of DL models in RA that automatically estimates joint inflammation, inheriting the preprocessing and model architecture from Chapter 3 and validation based on CAMs from Chapters 4 and 5. MRI scans of 2254 subjects from four study populations were divided into training, monitoring, test and validation sets for training and evaluation. This study demonstrates the ability of DL models to recognize the manually defined image biomarkers and to estimate severity as accurately as human experts.

In **Chapter 3**, we developed a series of DL models to detect early RA signs and predict which patients would develop RA within approximately two years after MRI scanning

(future RA development), and some other classification tasks. In this chapter, we first proposed a pre-processing method to select the most informative slices from 3D MRI scans with irregular sizes. Combined with a multi-input DL model architecture and a consistency-based loss function, this method overcomes the challenge of applying deep learning to RA-related datasets with a limited number of samples, complex and diverse anatomical, pathological structures and artefacts. The performance of the developed DL models in predicting future RA development is similar to that of human experts using statistical models with specific clinical variables. The results demonstrate not only the existence of image biomarkers for RA prediction, but also the ability of DL models to discover these image biomarkers, making it feasible for subsequent data-driven studies and corresponding chapters.

Chapter 4 provides a technical method for interpreting DL models. In this chapter, we resolved an intensity scaling problem in prevailing CAM algorithms, in which the normalization of CAMs is performed based only on a single input. This new method enables further accurate interpretation of DL models and additional validation of the reliability and explainability of these data-driven black boxes. Validated on eight different datasets with different modalities, this method provides an approach to visually check the consistency between the focus of DL models and expert knowledge at the individual level. It allows reader studies to validate the reliability and alignment with human experts of the trained DL models. Applying this method to the DL models in Chapter 3, we found that high-intensity signals from inflammatory signs were captured and highlighted by DL models, which is consistent with rheumatologists. Furthermore, in Chapter 4, we proposed a so-called "feature distinction" method that can determine the contribution of each feature extracted by DL models, to analyze both the datasets and DL models, extending the use of CAM algorithms. Through this feature distinction, we are able to monitor model overfitting, detect the difference between training and monitoring sets, locate confounders, and find principal features.

Based on Chapter 4, in **Chapter 5** we aimed to solve the problem of CAM that it is an individual-level analysis. Chapter 5 provides a segmentation-based framework that aggregates the information in the improved CAMs from Chapter 4 to draw population-level conclusions or generate new hypotheses. By directly comparing these conclusions or hypotheses with expert knowledge, it allows us to validate the reliability of DL models and discover the correlation between image biomarkers and DL model outputs. We validated this framework in a series of simulation experiments, demonstrating its accuracy and reliability, and then applied it to our RA prediction model to discover which regions in MRI scans are most informative for RA prediction. Under current experimental settings (datasets, data splitting, models and preprocessing method), tenosynovitis around flexor tendons contributes most to early RA detection and future RA development prediction.

Chapter 6 explores an additional application of the above methods - the discovery of new image biomarkers that could indicate early RA. In this chapter, a DL model first outputs a set of features, including the manually defined image biomarkers (Chapter 2), and then outputs RA labels (Chapter 3). By generating CAMs (Chapter 4) only for the paths that do not produce manually defined image biomarkers, the regions of potential new image biomarkers can be located by visualization and analyzed by aggregation (Chapter 5). We applied this method to the dataset that includes patients who developed RA after the MRI scanning, to check whether the method can find the image biomarkers relevant to RA. Since the lack of "undefined" "RA-relevant" image biomarkers, we took tenosynovitis out of the manually-defined image biomarkers and checked whether the approach would then consider regions of tenosynovitis as "new image biomarkers". These results indicated the feasibility of this fully data-driven method for discovering new image biomarkers under certain conditions.

7.2 Discussion on limitations and future work

7.2.1 Discussion per topic from each chapter

In this thesis, we developed a systematic framework for (1) automatic joint inflammation estimation; (2) fully data-driven future RA development prediction; (3) individual- and population-level interpretation of the DL model for RA prediction and joint inflammation estimation; and (4) discovering and locating potential new image biomarkers. This section further discusses the limitations of the methods proposed or developed in these chapters, and provides some potential solutions to these limitations.

- Limitation and future work of Chapter 2, automatic joint inflammation estimation.

In the automatic joint inflammation estimation, the major limitation is overestimation caused by overfitting. The models output all inflammation assessments at once for efficiency reasons, which may lead to overfitting of implicit correlations between specific inflamed regions unique in the used dataset. One phenomenon of this problem is that the performance of estimating total inflammation severity is higher than estimating inflammation severity for each anatomical structure. This phenomenon suggests that our automatic joint inflammation estimation works differently from human experts and potentially overfits this dataset – it may predict the inflammation severity of some joints based on implicit correlations among the joints. This phenomenon originates from the training strategy – the DL models were trained to produce joint inflammation assessments for all joints at the same time, in order to have higher training and inference efficiency. This training strategy increased the risk of taking shortcuts and learning from correlated inflammatory signs that are particularly prevalent in the specific dataset. This kind of speculation rather

than actual assessment could lead to over- or under-estimation on evaluating the performance of proposed DL models, where the high correlation of the DL models may bepartially obtained through correlation among some specific inflamed regions. This could be a serious problem to the model when the distribution and characteristics of a new dataset significantly differ from the training set - this kind of implicit correlation may not exist and the performance could drop substantially. A simple solution is to train one model for each joint, but this solution not only reduces efficiency (the cost of inference increases linearly with the number of joints), but also assumes that there should be no correlation among joints' inflammation. One of the potential future works could solve this problem with a more subtle strategy – training DL models with a customized augmentation method that can add inflammation to specific joints to balance the speculation and assessment.

- Limitation and future work of Chapter 3, fully data-driven future RA development prediction.

A limitation shared by Chapter 2 and Chapter 3 is the preprocessing method, and this limitation has a more substantial impact in Chapter 3. In the fully data-driven RA development prediction, the preprocessing, especially 2D slice selection, plays a substantial role in the performance of the whole method. The proposed automatic selection relies on the standard deviation of the signal intensities of each slice in the 3D MRI scans. This strategy originates from the observation that inflammation will increase the standard deviations within a 2D slice, therefore focusing on slices with higher standard deviations may help keep inflammation signals and filter out less informative slices. However, some other factors such as artifacts could also lead to high standard deviations, resulting in the wrong selection of 2D slices. By improving this preprocessing with more intelligent algorithms, more accurate DL models likely to be obtained by filtering out noise and preserving informative signals. Similarly, the augmentation in preprocessing can also be further improved. Augmentation substantially improves the performance of DL models applied to datasets with a limited number of samples, including the dataset in this research. Currently, the augmentation employed in this research is restricted to some basic augmentations such as rotation, spatial scaling, and translation. More advanced augmentation methods including simulation of MRI artefacts would probably further improve the model.

- Limitation and future work of Chapter 4 and 5, individual- and population-level interpretation of the DL model.

Feature analysis (Chapter 4) and aggregation (Chapter 5) for CAMs are general methods that can be applied to many fields, and we have investigated many major

technical details of these new approaches. However, two aspects remain unexplored – the definition of "importance" in CAMs and the quantitative influence of segmentation accuracy. We choose to use the mean activations of a region in this thesis as the definition of "importance" in CAMs, but other definitions may also be reasonable. The choice of "importance" requires quantitative and systematic investigation to figure out the optimal choice and the corresponding circumstances. Similarly, while the accuracy of segmentation could affect the accuracy of the CAM aggregation and its conclusion according to our intuition, we did not conduct a systematic and quantitative evaluation of how severe this influence is on the task of RA prediction. In the future, the influence of less accurate segmentation needs to be clarified when applying this data-driven hypothesis generator in practice.

- Limitation and future work of Chapter 6, discovering and locating potential new image biomarkers.

The last chapter of this thesis is a feasibility study on whether DL models and CAMs can serve as tools to discover new image biomarkers other than the defined features. While a series of simulation experiments were conducted to verify its reliability and accuracy, some limitations seriously restrict the generalizability of this framework and many open questions remain unsolved. The most serious limitation of this method is that both the prevalence and the "importance" (posterior probability of an output given this image biomarker) of the "undiscovered" new image biomarkers must reach a particular level and can then be learned by DL models. This limitation, derived from the information coupling between different nodes in neural networks, restricts this method to a tool that can only discover prevalent and significant image biomarkers. Personally, the author believes that a solution to this problem is to develop a series of technical methods to restrict the nodes within a DL model, decoupling the information and making the nodes' information as "pure" as possible. This is a potential and challenging future task that can extend the use of DL models, from model-based "learning" to model-based "teaching" .

7.2.2 General discussion

In addition to the discussion and future work of each chapter in detail, we would like to discuss some broader topics from both technical and clinical perspectives.

- Advanced DL techniques and classic image processing.

AI methods, especially DL methods, have developed rapidly in past thirteen years since the first time they were applied [165, 166, 167, 168, 169]. Since then, many advanced novel DL model architectures (e.g. fully convolutional networks[78], dense connections [170], attention mechanism [171], hybrid architectures [172], foundation

models [173]), training strategies (e.g. federation learning [174], self-supervision [83], multitasking [175], transfer learning [176]) were proposed and applied to different applications (e.g. segmentation [78], diagnosis [165], interaction [177], reconstruction [178], transformation [179]). In various senses, estimating joint inflammation, detecting early RA signs and predicting RA development should benefit from these methods and strategies. However, many attempts and pre-experiments in applying some of these advanced methods (e.g. dense connections [170], advanced architectures [94, 171], multitasking [175], transfer learning [176]) ended up failing to improve the performance due to methodological mismatch, data limitations, or task-specific constraints (and some of them cannot be applied because of the failure to meet the requirements or other reasons). These advanced methods usually have premises and are often designed for too specific tasks, therefore they may not contribute to this research. On the contrary, classic image processing methods contributed substantially to improving the performance of DL models (e.g. the preprocessing method based on histogram and morphology in Chapter 2 and 3). The author personally believes that this difference stems from two factors - the degree of abstraction and the way of application. Classic image processing methods are typically concrete to process a specific aspect of an image, they are also applied when the targeted problems occurred and were recognized. The entire process diverges when advanced DL methods (architectures and strategies) are applied, as they are data-driven to extract high-level abstract features. In current medical image processing, many problems cannot be easily solved without DL methods as they are too challenging for classic image processing. However, relying too much on deep learning methods can also lead to researches focusing on solving problems, rather than producing generalizable methods that extend our understanding and knowledge on the tasks. Therefore, it is a very interesting and valuable research direction to interpret these features from the data-driven feature extraction process of DL models to form new understanding and knowledge.

- Explainable DL methods.

For understanding the features extracted by DL models, explainable DL methods are emerging. Explainable DL methods are typically divided into interpretable model architectures and post-hoc analysis methods. Interpretable model architectures typically pose strict restrictions on the datasets and are therefore not widely employed. Post-hoc analysis includes methods such as saliency maps [58, 55, 57], feature importance [109, 113], filter or neuron visualization [180], error analysis and uncertainty estimation. In this thesis, saliency maps, CAM-based feature importance were applied, improved and extended to a population level. The explainability methods developed in this study are essentially simple and intuitive, such as global intensity scaling and aggregation across the dataset. Many studies in medical imaging applied

the saliency maps, assuming the problems (e.g. intensity scaling) were solved when these saliency maps were proposed. However, as the main DL research of images focused on non-medical images, the comparison between individual CAMs is typically not needed - as the population-level analysis. This reflects the different needs of different fields for the same technology, and finding such different interests becomes the source of innovation in this field.

- DL methods in the specific field of RA.

On the specific tasks in this RA-related study, the proposed DL models have achieved accuracy levels close to human experts [72, 10]. However, the generalizability and robustness requires further validation (as suggested in [181]), and these methods are therefore not contributing to clinical practice at present. This lack of generalizability originates from the difficulty of collecting large-scale MRI data in the narrow time window of early RA and accessing similar datasets [182] - a common difficulty in developing DL methods in medical imaging. Efforts are made to build foundation models [183, 184, 185] or use transfer learning [186] to alleviate this problem, yet the effects are frequently limited in both downstream tasks [187] and upstream segmentation [188]. Therefore, when other deep learning application fields are no longer (seriously) limited by data, the amount and quality of data remain to be the bottleneck of DL in medical image processing. To make matters worse, medical image data is not as easy to collect as natural images, natural language and other data - data collection is subject to many restrictions, such as the speed of disease progression, the prevalence of the disease and medical ethics. Since the collection of clinical data requires considerable time and cannot be simply accelerated, sharing data could be essential for further development of AI in the field of RA. Only by sharing data across different protocols, populations and time periods, a robust, generalizable and impactful automatic approach can be developed in this specific field of RA. However, this kind of data sharing also introduces ethical and information security risks. The dependency of DL methods on larger scale and more diverse data contradicts with the limited data capacity, individual diversity, ethical and data security risks in this field, making it challenging to conduct truly robust and impactful AI research for RA in a short term.

- The relationship between AI and clinicians.

Another topic is the relationship between AI and clinicians, which has been frequently discussed and questioned during the conferences the author attended. AI is typically considered to replace the work of humans and therefore creates job losses, even for medical practitioners [189]. However, in medical fields, AI can at best

serve as tools. On the one hand, the development and reliability of AI in medical fields is relatively slow compared to other fields, due to (1) limited capacity of high-quality and unbiased data, (2) medical ethics and responsibility [190], (3) unstructured data that could affect clinical decisions, (4) the dependency on standardized and predefined settings, (5) rapid update of medical knowledge [191]. On the other hand, psychological and social factors are such that most people, including AI developers, are actually reluctant to replace doctors with AI [192]. The main application scenarios of AI should be to replace repetitive operations to reduce the burden on doctors, and to provide assistance and reference when medical practitioners lack relevant medical information in certain scenarios. Take RA as an example, while preventing chronic RA requires early diagnosis, detecting potential RA requires long specialized training and is therefore not applicable for all doctors. Reachable AI tools can serve as tools to identify patients at risk and recommend transferring them to rheumatologists.

7.3 General conclusions

In summary, the studies in this thesis have established a basic framework for the prediction of RA development, assessment of joint inflammation, interpretation and validation of the DL models in these applications, and even the potential discovery of some knowledge using DL models. These studies provide a benchmark for the capabilities of DL models when applied to RA and joint inflammation in the future, with useful generalizable tools for validating and interpreting these DL models, which are valuable in many medical fields.

8

Samenvatting, discussie en toekomstig werk

In dit onderzoek hebben we een reeks modellen en methoden ontwikkeld voor het kwantificeren van ontstekingen, het voorspellen van toekomstige RA-ontwikkeling, het interpreteren van DL-modellen, het samenvoegen van informatie om populatieniveau-conclusies te trekken, en het combineren van al deze methoden om potentiële nieuwe beeldbiomarkers te ontdekken die bijdragen aan RA-voorspelling. In dit hoofdstuk vatten we kort de voorgaande hoofdstukken samen en bespreken we hun voordelen, beperkingen en mogelijke toekomstige onderzoeksrichtingen.

8.1 Samenvatting

In dit proefschrift biedt het eerste hoofdstuk een algemene introductie van het onderzoeksonderwerp, de motivatie en uitdagingen die overwonnen moeten worden, en de relatie tussen de volgende hoofdstukken. Het oorspronkelijke idee van het hele onderzoek was om een data-gedreven methode te ontwikkelen, onafhankelijk van expertkennis, om RA in een vroeg stadium te detecteren of voorspellen en ook om vroege tekenen van RA te vinden. Het hele idee is verdeeld in vier doelstellingen/stappen: (1) Hoofdstuk 2: Een potentiële toepassing van DL-modellen in de klinische praktijk, en de haalbaarheid van DL-modellen in het herkennen van handmatig gedefinieerde beeldbiomarkers - dit maakt de potentiële ontdekking van aanvullende beeldbiomarkers mogelijk naast deze vooraf gedefinieerde beeldbiomarkers; (2) Hoofdstuk 3: de haalbaarheid en de methode om toekomstige RA-ontwikkeling in MRI te voorspellen met DL-modellen; (3) Hoofdstukken 4 en 5: een algemeen kader om aan te tonen dat de DL-modellen betrouwbaar en verklaarbaar zijn, en of ze redeneren op basis van aanvullende beeldbiomarkers anders dan menselijke experts, op individueel en populatieniveau; (4) Hoofdstuk 6: Een haalbaar kader voor het ontdekken van nieuwe beeldbiomarkers op basis van CAM-algoritmen.

In **Hoofdstuk 2** onderzochten we een praktische toepassing van DL-modellen in RA die automatisch gewrichtsontsteking schat, waarbij de preprocessing en modelarchitectuur werden overgenomen uit Hoofdstuk 3 en validatie gebaseerd op CAMs uit Hoofdstukken 4 en 5. MRI-scans van 2254 proefpersonen uit vier onderzoekspopulaties werden verdeeld in trainings-, monitorings-, test- en validatiesets voor training en evaluatie. Deze studie toont het vermogen van DL-modellen aan om de handmatig

gedefinieerde beeldbiomarkers te herkennen en de ernst even goed in te schatten als menselijke experts.
In **Hoofdstuk 3** ontwikkelden we een reeks DL-modellen om vroege RA-tekenen te detecteren en te voorspellen welke patiënten RA zouden ontwikkelen binnen bijna twee jaar na de MRI-scan (toekomstige RA-ontwikkeling), en enkele andere classificatietaken. In dit hoofdstuk stelden we eerst een preprocessing voor om de meest informatieve slices te selecteren uit 3D MRI-scans met onregelmatige afmetingen. Gecombineerd met een multi-input DL-modelarchitectuur en een op consistentie gebaseerde verliesfunctie, overwint deze methode de uitdaging van het toepassen van deep learning op RA-gerelateerde datasets met een beperkt aantal samples, complexe en diverse anatomische, pathologische structuren en artefacten. De prestaties van de ontwikkelde DL-modellen in het voorspellen van toekomstige RA-ontwikkeling zijn vergelijkbaar met die van menselijke experts die statistische modellen gebruiken met specifieke klinische variabelen. De resultaten tonen niet alleen het bestaan van beeldbiomarkers voor RA-voorspelling aan, maar ook het vermogen van DL-modellen om deze beeldbiomarkers te ontdekken, wat het haalbaar maakt voor vervolgdata-gedreven studies en hoofdstukken.
Hoofdstuk 4 biedt een technische methode voor het interpreteren van DL-modellen. In dit hoofdstuk hebben we een intensiteitsschalingsprobleem opgelost in gangbare CAM-algoritmen, waarbij de normalisatie van CAMs wordt uitgevoerd op basis van een enkele input zelf. Deze nieuwe methode maakt een verdere nauwkeurige interpretatie van DL-modellen mogelijk en aanvullende validatie van de betrouwbaarheid en verklaarbaarheid van deze data-gedreven black boxes. Geverifieerd op acht verschillende datasets met verschillende modaliteiten, biedt deze methode een benadering om visueel de consistentie te controleren tussen de focus van DL-modellen en expertkennis op individueel niveau. Het maakt lezersstudies mogelijk om de betrouwbaarheid en afstemming met menselijke experts van de getrainde DL-modellen te valideren. Door deze methode toe te passen op de DL-modellen in Hoofdstuk 3, ontdekten we dat hoogintensiteitssignalen van ontstekingstekenen werden vastgelegd en benadrukt door DL-modellen, wat consistent is met reumatologen. Verder hebben we in Hoofdstuk 4 een zogenaamde "feature distinction"-methode voorgesteld die de bijdrage van elke door DL-modellen geëxtraheerde feature kan bepalen, om zowel de datasets als DL-modellen te analyseren, en het gebruik van CAM-algoritmen uit te breiden. Door deze feature distinction kunnen we modeloverfitting monitoren, het verschil tussen trainings- en monitoringssets detecteren, confounders lokaliseren en hoofdkenmerken vinden.
Gebaseerd op Hoofdstuk 4, hebben we in **Hoofdstuk 5** geprobeerd het probleem op te lossen dat CAM een analyse op individueel niveau is. Hoofdstuk 5 biedt een op segmentatie gebaseerd kader dat de informatie in de verbeterde CAMs uit Hoofdstuk

4 samenvoegt om populatieniveau-conclusies te trekken of nieuwe hypothesen te genereren. Door deze conclusies of hypothesen direct te vergelijken met expertkennis, kunnen we de betrouwbaarheid van DL-modellen valideren en de correlatie tussen beeldbiomarkers en DL-modeloutputs ontdekken. We hebben dit kader gevalideerd in een reeks simulatie-experimenten, waarbij we de nauwkeurigheid en betrouwbaarheid aantoonden, en pasten het vervolgens toe op ons RA-voorspellingsmodel om te ontdekken welke regio's in MRI-scans het meest informatief zijn voor RA-voorspelling. Onder de huidige experimentele instellingen (datasets, datasplitsing, modellen en preprocessing-methode) draagt tenosynovitis rond flexorpezen het meest bij aan vroege RA-detectie en toekomstige RA-ontwikkelingsvoorspelling.

Hoofdstuk 6 onderzoekt een aanvullende toepassing van de bovenstaande methoden - de ontdekking van nieuwe beeldbiomarkers die vroege RA kunnen aangeven. In dit hoofdstuk produceert een DL-model eerst een set features, inclusief de handmatig gedefinieerde beeldbiomarkers (Hoofdstuk 2), en vervolgens RA-labels (Hoofdstuk 3). Door CAMs te genereren (Hoofdstuk 4) alleen voor de paden die geen handmatig gedefinieerde beeldbiomarkers produceren, kunnen de regio's van potentiële nieuwe beeldbiomarkers worden gelokaliseerd door visualisatie en geanalyseerd door aggregatie (Hoofdstuk 5). We pasten deze methode toe op de dataset die patiënten omvat die RA ontwikkelden na de MRI-scan, om te controleren of de methode de beeldbiomarkers relevant voor RA kon vinden. Vanwege het ontbreken van "ongedefinieerde" "RA-relevante" beeldbiomarkers, hebben we tenosynovitis uit de handmatig gedefinieerde beeldbiomarkers gehaald en gecontroleerd of de aanpak de regio's van tenosynovitis zou benadrukken. Deze resultaten gaven de haalbaarheid aan van deze volledig data-gedreven methode voor het ontdekken van nieuwe beeldbiomarkers onder bepaalde omstandigheden.

8.2 Discussie over beperkingen en toekomstig werk

8.2.1 Discussie binnen elk hoofdstuk

In de automatische schatting van gewrichtsontsteking is de prestatie van het schatten van de totale ontstekingsernst hoger dan het schatten van de ontstekingsernst voor elke anatomische structuur. Dit fenomeen suggereert dat onze automatische gewrichtsontstekingsschatting anders werkt dan menselijke experts en mogelijk overfit op deze dataset – het kan de ontstekingsernst van sommige gewrichten voorspellen op basis van impliciete correlaties tussen de gewrichten. Deze feature komt voort uit de trainingsstrategie – de DL-modellen werden getraind om de gewrichtsontstekings-beoordeling voor alle gewrichten tegelijkertijd te produceren, om hogere trainings- en inferentie-efficiëntie te hebben. Deze trainingsstrategie verhoogde het risico op het nemen van shortcuts en het leren van gecorreleerde ontstekingstekenen die bijzonder prevalent zijn in de specifieke dataset. Dit soort speculatie in plaats van daadwerkelijke

beoordeling kan leiden tot over- of onderschatting bij het evalueren van de prestaties van de voorgestelde DL-modellen, waarbij de hoge correlatie van de DL-modellen gedeeltelijk kan worden verkregen door correlatie tussen sommige specifieke ontstoken regio's. Dit kan een ernstig probleem zijn voor het model wanneer de distributie en kenmerken van een nieuwe dataset aanzienlijk verschillen van de trainingsset – deze soort impliciete correlatie bestaat mogelijk niet en de prestaties kunnen aanzienlijk dalen. Een eenvoudige oplossing is om één model per gewricht te trainen, maar deze oplossing vermindert niet alleen de efficiëntie (de kosten van inferentie nemen lineair toe met het aantal gewrichten), maar gaat er ook van uit dat er geen correlatie mag zijn tussen gewrichtsontstekingen. Een van de potentiële toekomstige werken zou dit probleem kunnen oplossen met een subtielere strategie – het trainen van DL-modellen met een aangepaste augmentatiemethode die ontsteking aan specifieke gewrichten kan toevoegen om de speculatie en beoordeling in balans te brengen.

In de volledig data-gedreven RA-ontwikkelingsvoorspelling speelt de preprocessing, met name 2D slice-selectie, een substantiële rol in de prestaties van de hele methode. De voorgestelde automatische selectie is gebaseerd op de standaarddeviatie van de signaalintensiteiten van elke slice in de 3D MRI-scans. Deze strategie komt voort uit de observatie dat ontsteking de standaarddeviaties binnen een 2D slice zal verhogen, daarom kan het focussen op slices met hogere standaarddeviaties helpen om ontstekingssignalen te behouden en minder informatieve slices te filteren. Echter, sommige andere factoren zoals artefacten kunnen ook leiden tot hoge standaarddeviaties, wat resulteert in de verkeerde selectie van 2D slices. Door deze preprocessing te verbeteren met intelligentere algoritmen, kunnen waarschijnlijk nauwkeurigere DL-modellen worden verkregen door ruis te filteren en informatieve informatie te behouden. Evenzo kan de augmentatie in preprocessing ook verder worden verbeterd. Augmentatie verbetert de prestaties van DL-modellen aanzienlijk wanneer ze worden toegepast op datasets met een beperkt aantal samples, inclusief de dataset in dit onderzoek. Momenteel is de augmentatie die in dit onderzoek wordt gebruikt beperkt tot enkele basisaugmentaties zoals rotatie, spatial scaling en translatie. Meer geavanceerde augmentatiemethoden, inclusief simulatie van MRI-artefacten, zouden waarschijnlijk het model verder verbeteren.

Feature-analyse (Hoofdstuk 4) en aggregatie (Hoofdstuk 5) voor CAMs zijn algemene methoden die op veel velden kunnen worden toegepast, en we hebben veel belangrijke technische details van deze nieuwe benaderingen onderzocht. Echter, twee aspecten blijven ononderzocht – de definitie van "belangrijkheid" in CAMs en de kwantitatieve invloed van segmentatienauwkeurigheid. We kozen ervoor om de gemiddelde activaties van een regio in dit proefschrift te gebruiken als de definitie van "belangrijkheid" in CAMs, maar andere definities kunnen ook redelijk zijn. De keuze van "belangrijkheid" vereist kwantitatief en systematisch onderzoek om de

optimale keuze en de bijbehorende omstandigheden te achterhalen. Evenzo, hoewel de nauwkeurigheid van segmentatie de nauwkeurigheid van de CAM-aggregatie en zijn conclusie volgens onze intuïtie zou kunnen beïnvloeden, hebben we geen systematische en kwantitatieve evaluatie uitgevoerd van hoe ernstig deze invloed is op de taak van RA-voorspelling. In de toekomst moet de invloed van minder nauwkeurige segmentatie worden opgehelderd wanneer deze data-gedreven hypothesegenerator in de praktijk wordt toegepast.

Het laatste hoofdstuk van dit proefschrift is een haalbaarheidsstudie over of DL-modellen en CAMs kunnen dienen als tools om nieuwe beeldbiomarkers te ontdekken naast de gedefinieerde features. Hoewel een reeks simulatie-experimenten werd uitgevoerd om de betrouwbaarheid en nauwkeurigheid te verifiëren, beperken enkele beperkingen de generaliseerbaarheid van dit kader ernstig en blijven veel open vragen onopgelost. De meest ernstige beperking van deze methode is dat zowel de prevalentie als de "belangrijkheid" (posterior waarschijnlijkheid van een output gegeven deze beeldbiomarker) van de "onontdekte" nieuwe beeldbiomarkers een bepaald niveau moeten bereiken en vervolgens kunnen worden geleerd door DL-modellen. Deze beperking, afgeleid van de informatiekoppeling tussen verschillende nodes in neurale netwerken, beperkt deze methode tot een tool die alleen prevalente en significante beeldbiomarkers kan ontdekken. Persoonlijk gelooft de auteur dat een oplossing voor dit probleem is om een reeks technische methoden te ontwikkelen om de nodes binnen een DL-model te beperken, de informatie te ontkoppelen en de informatie van de nodes zo "puur" mogelijk te maken. Dit is een potentiële en uitdagende toekomstige taak die het gebruik van DL-modellen kan uitbreiden, van modelgebaseerd "leren" naar modelgebaseerd "onderwijzen".

8.2.2 Algemene discussie

Naast de gedetailleerde discussie en toekomstig werk van elk hoofdstuk, willen we enkele bredere onderwerpen bespreken vanuit zowel technisch als klinisch perspectief.

- Geavanceerde DL-technieken en klassieke beeldverwerking.
- Verklaarbare DL-methoden.
- DL-methoden in het specifieke veld van RA.
- De relatie tussen AI en clinici.

AI-methoden, met name DL-methoden, hebben zich de afgelopen dertien jaar snel ontwikkeld sinds ze voor het eerst werden toegepast [165, 166, 167, 168, 169]. Sindsdien zijn veel geavanceerde nieuwe DL-modelarchitecturen (zoals volledig convolutionele netwerken [78], dense connections [170], attention mechanism [171], hybride architecturen [172], foundation models [173]), trainingsstrategieën (zoals

federated learning [174], self-supervision [83], multitasking [175], transfer learning [176]) voorgesteld en toegepast op verschillende toepassingen (zoals segmentatie [78], diagnose [165], interactie [177], reconstructie [178], transformatie [179]). In verschillende opzichten kunnen het schatten van gewrichtsontsteking, het detecteren van vroege RA-signalen en het voorspellen van RA-ontwikkeling baat hebben bij deze methoden en strategieën. Echter, veel pogingen en vooronderzoeken bij het toepassen van sommige van deze geavanceerde methoden (zoals dense connections [170], geavanceerde architecturen [94, 171], multitasking [175], transfer learning [176]) resulteerden in het niet verbeteren van de prestaties vanwege methodologische mismatch, databeperkingen of taakspecifieke beperkingen (en sommige konden niet worden toegepast vanwege het niet voldoen aan vereisten of andere redenen). Deze geavanceerde methoden hebben meestal voorwaarden en zijn vaak ontworpen voor specifieke taken, waardoor ze mogelijk niet bijdragen aan dit onderzoek. In tegenstelling hiermee hebben klassieke beeldverwerkingsmethoden aanzienlijk bijgedragen aan het verbeteren van de prestaties van DL-modellen (zoals de preprocessing-methode gebaseerd op histogram en morfologie in Hoofdstuk 2 en 3). De auteur is persoonlijk van mening dat dit verschil voortkomt uit twee factoren - de mate van abstractie en de wijze van toepassing. Klassieke beeldverwerkingsmethoden zijn typisch concreet om een specifiek aspect van een beeld te verwerken, ze worden ook toegepast wanneer de beoogde problemen zich voordoen en worden herkend. Het hele proces verloopt anders wanneer geavanceerde DL-methoden (architecturen en strategieën) worden toegepast, aangezien deze data-gedreven zijn om hoogwaardige abstracte kenmerken te extraheren. In de huidige medische beeldverwerking kunnen veel problemen niet eenvoudig worden opgelost zonder DL-methoden, omdat ze uitdagend zijn voor klassieke beeldverwerking. Echter, overmatige afhankelijkheid van deep learning-methoden kan er ook toe leiden dat onderzoeken zich richten op het oplossen van problemen, in plaats van het produceren van generaliseerbare methoden die ons begrip en kennis van de taken uitbreiden. Daarom is het een zeer interessante en waardevolle onderzoeksrichting om deze kenmerken te interpreteren vanuit het data-gedreven kenmerkextractieproces van DL-modellen om nieuw begrip en kennis te vormen.

Voor het begrijpen van de door DL-modellen geëxtraheerde kenmerken, komen verklaarbare DL-methoden op. Verklaarbare DL-methoden worden typisch onderverdeeld in interpreteerbare modelarchitecturen en post-hoc-analysemethoden. Interpreteerbare modelarchitecturen leggen meestal strikte beperkingen op aan de datasets en worden daarom niet veel gebruikt in veel DL-gebaseerde benaderingen. Post-hoc-analysemethoden omvatten saliency maps [58, 55, 57], feature importance [109, 113], filter- of neuronvisualisatie [180], foutenanalyse en onzekerheidsschatting. In deze studie werden saliency maps, CAM-gebaseerde feature importance toegepast, verbeterd en uitgebreid naar populatieniveau. De verklaarbaarheidsmethoden die

in deze studie zijn ontwikkeld, zijn in wezen eenvoudig en intuïtief, zoals globale intensiteitsschaal en aggregatie over de dataset. Veel studies in het medische beeldveld pasten ook deze saliency maps toe, in de veronderstelling dat de problemen (zoals intensiteitsschaal) waren opgelost toen deze saliency maps werden voorgesteld. De belangrijkste reden waarom deze problemen tot op de dag van vandaag voortduren, kan zijn dat - terwijl andere velden ook populatieniveau-analyse kunnen vereisen, het medische veld unieke nadruk legt op cross-individuele vergelijking vanwege klinische variabiliteit en diagnostische betrouwbaarheid. Dit weerspiegelt de verschillende behoeften van verschillende velden voor dezelfde technologie, en het vinden van dergelijke verschillende belangen wordt de bron van innovatie in dit veld.

Voor de specifieke taken in deze RA-gerelateerde studie hebben de voorgestelde DL-modellen voor RA-gerelateerde taken niveaus bereikt die dicht bij menselijke experts liggen [72, 10]. Echter, de generaliseerbaarheid en robuustheid vereisen verdere validatie (zoals gesuggereerd in [181]), en deze methoden dragen daarom momenteel niet bij aan de klinische praktijk. Dit gebrek aan generaliseerbaarheid komt voort uit de moeilijkheid om grootschalige MRI-gegevens te verzamelen in het smalle tijdvenster van vroege RA en vergelijkbare datasets te benaderen [182] - een veelvoorkomende moeilijkheid bij het ontwikkelen van DL-methoden in medische beeldvelden. Er worden inspanningen geleverd om foundation models [183, 184, 185] te bouwen of transfer learning [186] te gebruiken om dit probleem te verlichten, maar de effecten zijn vaak beperkt in zowel downstream-taken [187] als upstream-segmentatie [188]. Daarom, wanneer andere deep learning-toepassingsvelden niet langer (ernstig) worden beperkt door data, blijven de hoeveelheid en kwaliteit van gegevens het knelpunt van DL in medische beeldverwerking. Tot overmaat van ramp zijn medische beeldgegevens niet zo gemakkelijk te vergroten als natuurlijke beelden, natuurlijke taal en andere gegevens - gegevensverzameling is onderworpen aan veel beperkingen, zoals de snelheid van ziekteprogressie, de prevalentie van de ziekte en medische ethiek. Omdat het verzamelen van klinische gegevens aanzienlijke tijd vergt en niet eenvoudig kan worden versneld, kan het delen van gegevens essentieel zijn voor verdere ontwikkeling van AI op het gebied van RA. Alleen door gegevens te delen over verschillende protocollen, populaties en tijdsperioden heen, kan een robuuste, generaliseerbare en impactvolle automatische aanpak worden ontwikkeld in dit specifieke veld van RA. Dit soort gegevensdeling brengt echter ook ethische en informatiebeveiligingsrisico's met zich mee. De afhankelijkheid van DL-methoden van grotere schaal en meer diverse gegevens staat haaks op de beperkte gegevenscapaciteit, individuele diversiteit, ethische en gegevensbeveiligingsrisico's in dit veld, waardoor het uitdagend is om op korte termijn echt robuust en impactvol AI-onderzoek voor RA uit te voeren.

Een ander onderwerp is de relatie tussen AI en clinici, die tijdens de conferenties

die de auteur heeft bijgewoond vaak is besproken en in twijfel getrokken. AI wordt doorgaans beschouwd als een vervanging voor het werk van mensen en veroorzaakt daarom banenverlies, zelfs voor medische professionals [189]. In medische velden kan AI echter hoogstens als hulpmiddelen dienen. Enerzijds is de ontwikkeling en betrouwbaarheid van AI in medische velden relatief traag in vergelijking met andere velden, vanwege (1) beperkte capaciteit van hoogwaardige en onbevooroordeelde gegevens, (2) medische ethiek en verantwoordelijkheid [190], (3) ongestructureerde gegevens die klinische beslissingen kunnen beïnvloeden, (4) de afhankelijkheid van gestandaardiseerde en vooraf gedefinieerde omstandigheden, (5) snelle update van medische kennis [191]. Anderzijds zorgen psychologische en sociale factoren ervoor dat de meeste mensen, inclusief AI-ontwikkelaars, eigenlijk terughoudend zijn om artsen te vervangen door AI [192]. De belangrijkste toepassingsscenario's van AI moeten zijn om repetitieve handelingen te vervangen om de last voor artsen te verminderen, en om hulp en referentie te bieden wanneer medische professionals in bepaalde scenario's over onvoldoende relevante medische capaciteiten beschikken. Neem RA als voorbeeld: hoewel het voorkomen van chronische RA vroege diagnose vereist, vereist het detecteren van potentiële RA veel gespecialiseerde training en is daarom niet toepasbaar voor alle artsen. Bereikbare AI-hulpmiddelen kunnen dienen als hulpmiddelen om potentiële patiënten in de menigte te identificeren en aan te bevelen om ze door te verwijzen naar professionele reumatologen.

8.3 Algemene conclusies

Samenvattend hebben de studies in dit proefschrift een basiskader gevestigd voor de voorspelling van RA-ontwikkeling, beoordeling van gewrichtsontsteking, interpretatie en validatie van de DL-modellen in deze toepassingen, en zelfs de potentiële ontdekking van enkele kennis met behulp van DL-modellen. Deze studies bieden een benchmark voor de mogelijkheden van DL-modellen wanneer ze in de toekomst worden toegepast op RA en gewrichtsontsteking, met nuttige generaliseerbare hulpmiddelen voor het valideren en interpreteren van deze DL-modellen, wat waardevol is in veel medische velden.

References

[1] E. Aizenberg. “Computer-aided techniques for assessment of MRI-detected inflammation for early identification of inflammatory arthritis”. PhD thesis. Leiden University, 2019.

[2] NIH. *Illustration of joint affected by rheumatoid arthritis*. `https://commons.wikimedia.org/wiki/File:Rheumatoid_arthritis_joint.gif`. [Internet]. 2007.

[3] V. Majithia and S. A. Geraci. “Rheumatoid Arthritis: Diagnosis and Management”. In: *American Journal of Medicine* 120 (11 2007), pages 936–939.

[4] N. Conrad, S. Misra, J. Y. Verbakel, et al. “Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK”. In: *The Lancet* 401.10391 (2023), pages 1878–1890.

[5] Y. Gao, Y. Zhang, and X. Liu. “Rheumatoid arthritis: pathogenesis and therapeutic advances”. In: *MedComm* 5.3 (2024), e509.

[6] P. Brown, A. G. Pratt, and K. L. Hyrich. “Therapeutic advances in rheumatoid arthritis”. In: *Bmj* 384 (2024).

[7] M. P. V. D. Linden, S. L. Cessie, K. Raza, et al. “Long-term impact of delay in assessment of patients with early arthritis”. In: *Arthritis and Rheumatism* 62 (12 Dec. 2010), pages 3537–3546.

[8] B. Heidari. “Rheumatoid Arthritis: Early diagnosis and treatment outcomes”. In: *Caspian journal of internal medicine* 2.1 (2011), page 161.

[9] J. F. Baker, J. O’ Dell, and P. Seo. “Diagnosis and differential diagnosis of rheumatoid arthritis”. In: *UpToDate. Published July* 28 (2023).

[10] X. M. Matthijssen, F. Wouters, D. M. Boeters, et al. “A search to the target tissue in which RA-specific inflammation starts: a detailed MRI study to improve identification of RA-specific features in the phase of clinically suspect arthralgia”. In: *Arthritis research & therapy* 21 (2019), pages 1–11.

[11] D. Shamonin, Y. LI, T. Hassanzadeh, et al. “POS0920 QUANTIFICATION OF TENOSYNOVITIS IN RA FROM WRIST MRIs, BASED ON DEEP LEARNING”. In: *Annals of the Rheumatic Diseases* 82.Suppl 1 (2023), pages 770–771. eprint: `https://ard.bmj.com/content/82/Suppl_1/770.2.full.pdf`.

[12] H. W. van Steenbergen, D. Aletaha, L. J. Beaart-van de Voorde, et al. “EULAR definition of arthralgia suspicious for progression to rheumatoid arthritis”. In: *Annals of the rheumatic diseases* 76.3 (2017), pages 491–496.

[13] S. J. Khidir, P. H. de Jong, A. Willemze, et al. “Clinically suspect arthralgia and rheumatoid arthritis: patients’ perceptions of illness”. In: *Joint Bone Spine* 91.6 (2024), page 105751.

[14] K. L. Moore and A. F. Dalley. *Clinically oriented anatomy*. Wolters kluwer india Pvt Ltd, 2018.

[15] S. Standring, H. Ellis, J. Healy, et al. “Gray’s anatomy: the anatomical basis of clinical practice”. In: *American journal of neuroradiology* 26.10 (2005), page 2703.

[16] R. Drake, A. W. Vogl, and A. W. Mitchell. *Gray’s anatomy for students E-book*. Elsevier Health Sciences, 2009.

[17] R. Ten Brinck, H. Van Steenbergen, and A. van der Helm–van Mil. “Sequence of joint tissue inflammation during rheumatoid arthritis development”. In: *Arthritis research & therapy* 20 (2018), pages 1–8.

[18] E. Olech, J. V. Crues, D. E. Yocum, and J. T. Merrill. “Bone marrow edema is the most specific finding for rheumatoid arthritis (RA) on noncontrast magnetic resonance imaging of the hands and wrists: a comparison of patients with RA and healthy controls”. In: *The Journal of rheumatology* 37.2 (2010), pages 265–274.

[19] D. I. Krijbolder, X. M. Matthijssen, B. T. van Dijk, et al. “The natural sequence in which subclinical inflamed joint tissues subside or progress to rheumatoid arthritis: a study of serial MRIs in the TREAT EARLIER trial”. In: *Arthritis & Rheumatology* 75.9 (2023), pages 1512–1521.

[20] Y. J. Dakkak, B. T. van Dijk, F. P. Jansen, et al. “Evidence for the presence of synovial sheaths surrounding the extensor tendons at the metacarpophalangeal joints: a microscopy study”. In: *Arthritis Research & Therapy* 24.1 (2022), page 154.

[21] D. A. Ton, B. T. van Dijk, H. W. van Steenbergen, A. H. van der Helm-van, et al. “Forefoot inflammation in recent-onset ACPA-positive and ACPA-negative RA: clinically similar, but different in underlying inflamed tissues”. In: *RMD open* 10.4 (2024), e004722.

[22] M. Østergaard, C. Peterfy, P. Conaghan, et al. “OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system.” In: *The Journal of Rheumatology* 30.6 (2003), pages 1385–1386. eprint: `https://www.jrheum.org/content/30/6/1385.full.pdf`.

[23] E. A. Haavardsholm, M. Østergaard, B. J. Ejbjerg, et al. “Introduction of a novel magnetic resonance imaging tenosynovitis score for rheumatoid arthritis: reliability in a multireader longitudinal study”. In: *Annals of the rheumatic diseases* 66.9 (2007), pages 1216–1220.

[24] Y. K. Tan and P. G. Conaghan. “Imaging in rheumatoid arthritis”. In: *Best Practice & Research Clinical Rheumatology* 25.4 (2011). Established Rheumatoid Arthritis, pages 569–584.

[25] E. Aizenberg, D. P. Shamonin, M. Reijnierse, et al. “Automatic quantification of tenosynovitis on MRI of the wrist in patients with early arthritis: a feasibility study”. In: *European Radiology* 29 (8 Aug. 2019), pages 4477–4484.

[26] F. McQueen, V. Beckley, J. Crabbe, et al. “Magnetic resonance imaging evidence of tendinopathy in early rheumatoid arthritis predicts tendon rupture at six years”. In: *Arthritis & Rheumatism* 52.3 (2005), pages 744–751.

[27] R. J. Wakefield, P. V. Balint, M. Szkudlarek, et al. “Musculoskeletal ultrasound including definitions for ultrasonographic pathology.” In: *The Journal of rheumatology* 32.12 (2005), pages 2485–2487.

[28] Y. K. Tan, M. Østergaard, and P. G. Conaghan. “Imaging tools in rheumatoid arthritis: ultrasound vs magnetic resonance imaging”. In: *Rheumatology* 51.suppl_7 (2012), pages vii36–vii42.

[29] W. P. Maksymowych. “The role of imaging in the diagnosis and management of axial spondyloarthritis”. In: *Nature Reviews Rheumatology* 15.11 (2019), pages 657–672.

[30] E. Naredo, I. Möller, A. Cruz, et al. “Power Doppler ultrasonographic monitoring of response to anti–tumor necrosis factor therapy in patients with rheumatoid arthritis”. In: *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 58.8 (2008), pages 2248–2256.

[31] M. Østergaard, C. G. Peterfy, P. Bird, et al. “The OMERACT rheumatoid arthritis magnetic resonance imaging (MRI) scoring system: updated recommendations by the OMERACT MRI in arthritis working group”. In: *The Journal of rheumatology* 44.11 (2017), pages 1706–1712.

[32] C. Yiu, J. F. Griffith, F. Xiao, et al. “Automated quantification of wrist bone marrow oedema, pre-and post-treatment, in early rheumatoid arthritis”. In: *Rheumatology Advances in Practice* 8.3 (2024), rkae073.

[33] Y. Mao, K. Imahori, W. Fang, et al. “Artificial Intelligence Quantification of Enhanced Synovium Throughout the Entire Hand in Rheumatoid Arthritis on Dynamic Contrast-Enhanced MRI”. In: *Journal of Magnetic Resonance Imaging* (2024).

[34] Y. Kobayashi, T. Kamishima, H. Sugimori, et al. “Quantification of hand synovitis in rheumatoid arthritis: Arterial mask subtraction reinforced with mutual information can improve accuracy of pixel-by-pixel time–intensity curve shape analysis in dynamic MRI”. In: *Journal of magnetic resonance imaging* 48.3 (2018), pages 687–694.

[35] A. S. Chand, A. McHaffie, A. W. Clarke, et al. “Quantifying synovitis in rheumatoid arthritis using computer-assisted manual segmentation with 3 tesla MRI scanning”. In: *Journal of Magnetic Resonance Imaging* 33.5 (2011), pages 1106–1113.

[36] E. Aizenberg, E. A. Roex, W. P. Nieuwenhuis, et al. “Automatic quantification of bone marrow edema on MRI of the wrist in patients with early arthritis: A feasibility study”. In: *Magnetic Resonance in Medicine* 79 (2 Feb. 2018), pages 1127–1134.

[37] D. Shamonin, Y. Li, T. Hassanzadeh, et al. “OP0190 QUANTIFICATION OF TENOSYNOVITIS, SYNOVITIS AND BONE MARROW EDEMA IN RHEUMATOID ARTHRITIS FROM WRIST MRIs, BASED ON DEEP LEARNING”. In: *Annals of the Rheumatic Diseases* 83.Suppl 1 (2024), pages 134–135. eprint: `https://ard.bmj.com/content/83/Suppl_1/134.full.pdf`.

[38] K. Nagpal, D. Foote, Y. Liu, et al. “Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer”. In: *NPJ digital medicine* 2.1 (2019), page 48.

[39] A. H. M. Linkon, M. M. Labib, T. Hasan, M. Hossain, et al. “Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study”. In: *Informatics in Medicine Unlocked* 24 (2021), page 100582.

[40] J. Rohrbach, T. Reinhard, B. Sick, and O. Dürr. “Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks”. In: *Computers & Electrical Engineering* 78 (2019), pages 472–481.

[41] R. W. Stidham, W. Liu, S. Bishu, et al. “Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis”. In: *JAMA network open* 2.5 (2019), e193963–e193963.

[42] G. Luo, S. Dong, W. Wang, et al. “Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification”. In: *Medical image analysis* 59 (2020), page 101591.

[43] M. W. Brejnebøl, P. Hansen, J. U. Nybing, et al. “External validation of an artificial intelligence tool for radiographic knee osteoarthritis severity classification”. In: *European Journal of Radiology* 150 (2022), page 110249.

[44] W. Bulten, H. Pinckaers, H. van Boven, et al. “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study”. In: *The Lancet Oncology* 21.2 (2020), pages 233–241.

[45] J. K. H. Andersen, J. S. Pedersen, M. S. Laursen, et al. “Neural networks for automatic scoring of arthritis disease activity on ultrasound images”. In: *RMD open* 5.1 (2019), e000891.

[46] T. Araújo, G. Aresta, L. Mendonça, et al. “DR| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images”. In: *Medical Image Analysis* 63 (2020), page 101715.

[47] J. D. Fauw, J. R. Ledsam, B. Romera-Paredes, et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: *Nature Medicine* 24 (9 Sept. 2018), pages 1342–1350.

[48] A. Esteva, B. Kuprel, R. A. Novoa, et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542 (7639 Feb. 2017), pages 115–118.

[49] V. Gulshan, L. Peng, M. Coram, et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *JAMA - Journal of the American Medical Association* 316 (22 Dec. 2016), pages 2402–2410.

[50] X. Li, H. Xiong, X. Li, et al. “Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond”. In: *Knowledge and Information Systems* 64.12 (2022), pages 3197–3234.

[51] P. W. Koh and P. Liang. “Understanding Black-box Predictions via Influence Functions”. In: *Proceedings of the 34th International Conference on Machine Learning*. Edited by D. Precup and Y. W. Teh. Volume 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pages 1885–1894.

[52] R. Baldock, H. Maennel, and B. Neyshabur. “Deep Learning Through the Lens of Example Difficulty”. In: *Advances in Neural Information Processing Systems*. Edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, et al. Volume 34. Curran Associates, Inc., 2021, pages 10876–10889.

[53] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. “Identifying Mislabeled Data using the Area Under the Margin Ranking”. In: *Advances in Neural Information Processing Systems*. Edited by H. Larochelle, M. Ranzato, R. Hadsell, et al. Volume 33. Curran Associates, Inc., 2020, pages 17044–17056.

[54] Q. Zhang, Y. N. Wu, and S.-C. Zhu. “Interpretable convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 8827–8836.

[55] B. Zhou, A. Khosla, A. Lapedriza, et al. “Learning Deep Features for Discriminative Localization”. In: (Dec. 2015).

[56] Y. Goyal, Z. Wu, J. Ernst, et al. “Counterfactual Visual Explanations”. In: *Proceedings of the 36th International Conference on Machine Learning*. Edited by K. Chaudhuri and R. Salakhutdinov. Volume 97. Proceedings of Machine Learning Research. PMLR, June 2019, pages 2376–2384.

[57] Y. Li, T. Hassanzadeh, D. P. Shamonin, et al. “Integrated feature analysis for deep learning interpretation and class activation maps”. In: *arXiv preprint arXiv:2407.01142* (2024).

[58] R. R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: (Oct. 2016).

[59] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pages 839–847.

[60] H. W. van Steenbergen, L. Mangnus, M. Reijnierse, et al. “Clinical factors, anticitrullinated peptide antibodies and MRI-detected subclinical inflammation in relation to progression from clinically suspect arthralgia to arthritis”. In: *Annals of the rheumatic diseases* 75.10 (2016), pages 1824–1830.

[61] A. Kleyer, M. Krieter, I. Oliveira, et al. “High prevalence of tenosynovial inflammation before onset of rheumatoid arthritis and its link to progression to RA—A combined MRI/CT study”. In: *Seminars in arthritis and rheumatism*. Volume 46. 2. Elsevier. 2016, pages 143–150.

[62] W. P. Nieuwenhuis, H. W. van Steenbergen, L. Mangnus, et al. “Evaluation of the diagnostic accuracy of hand and foot MRI for early rheumatoid arthritis”. In: *Rheumatology* 56.8 (2017), pages 1367–1377.

[63] S. Ajeganova and T. Huizinga. “Sustained remission in rheumatoid arthritis: latest evidence and clinical considerations”. In: *Therapeutic advances in musculoskeletal disease* 9.10 (2017), pages 249–262.

[64] D. I. Krijbolder, M. Verstappen, B. T. van Dijk, et al. “Intervention with methotrexate in patients with arthralgia at risk of rheumatoid arthritis to reduce the development of persistent arthritis and its disease burden (TREAT EARLIER): a randomised, double-blind, placebo-controlled, proof-of-concept trial”. In: *The Lancet* 400.10348 (2022), pages 283–294.

[65] C. Cano-Espinosa, G. González, G. R. Washko, et al. “Automated Agatston score computation in non-ECG gated CT scans using deep learning”. In: *Proceedings of SPIE–the International Society for Optical Engineering*. Volume 10574. NIH Public Access. 2018.

[66] B. D. De Vos, J. M. Wolterink, T. Leiner, et al. “Direct automatic coronary calcium scoring in cardiac and chest CT”. In: *IEEE transactions on medical imaging* 38.9 (2019), pages 2127–2138.

[67] D. Mu, J. Bai, W. Chen, et al. “Calcium scoring at coronary CT angiography using deep learning”. In: *Radiology* 302.2 (2022), pages 309–316.

[68] J. Jia, I. Hernández-Girón, A. A. Schouffoer, et al. “Explainable fully automated CT scoring of interstitial lung disease for patients suspected of systemic sclerosis by cascaded regression neural networks and its comparison with experts”. In: *Scientific Reports* 14.1 (2024), page 26666.

[69] B. Astuto, I. Flament, N. K. Namiri, et al. “Automatic deep learning–assisted detection and grading of abnormalities in knee MRI studies”. In: *Radiology: Artificial Intelligence* 3.3 (2021), e200165.

[70] P. Chen, L. Gao, X. Shi, et al. “Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss”. In: *Computerized Medical Imaging and Graphics* 75 (2019), pages 84–92.

[71] M. Schlereth, M. Y. Mutlu, J. Utz, et al. “Deep learning-based classification of erosion, synovitis and osteitis in hand MRI of patients with inflammatory arthritis”. In: *RMD open* 10.2 (2024), e004273.

[72] Y. Li, T. Hassanzadeh, D. P. Shamonin, et al. “Rheumatoid arthritis classification and prediction by consistency-based deep learning using extremity MRI scans”. In: *Biomedical Signal Processing and Control* 91 (2024), page 105990.

[73] P. Maragos. “Differential morphology and image processing”. In: *IEEE Transactions on Image Processing* 5 (6 June 1996), pages 922–937.

[74] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18 (2 Feb. 2021), pages 203–211.

[75] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762.

[76] A. Jaegle, S. Borgeaud, J. Alayrac, et al. “Perceiver IO: A General Architecture for Structured Inputs & Outputs”. In: *CoRR* abs/2107.14795 (2021). arXiv: 2107.14795.

[77] K. He, X. Zhang, S. Ren, and J. Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *CoRR* abs/1502.01852 (2015). arXiv: 1502.01852.

[78] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: (May 2015).

[79] A. A. Tolpadi, J. Luitjens, F. G. Gassert, et al. “Synthetic inflammation imaging with PatchGAN deep learning networks”. In: *Bioengineering* 10.5 (2023), page 516.

[80] M. L. Hetland, B. Ejbjerg, K. Hørslev-Petersen, et al. “MRI bone oedema is the strongest predictor of subsequent radiographic progression in early rheumatoid arthritis. Results from a 2-year randomised controlled trial (CIMESTRA)”. In: *Annals of the Rheumatic Diseases* 68 (3 Mar. 2009), pages 384–390.

[81] F. Xiao, J. F. Griffith, A. L. Hilkens, et al. “ERAMRS: a new MR scoring system for early rheumatoid arthritis of the wrist”. In: *European Radiology* 29 (10 Oct. 2019), pages 5646–5654.

[82] P. Bøyesen, E. A. Haavardsholm, M. Østergaard, et al. “MRI in early rheumatoid arthritis: Synovitis and bone marrow oedema are independent predictors of subsequent radiographic progression”. In: *Annals of the Rheumatic Diseases* 70 (3 Mar. 2011), pages 428–433.

[83] K. He, X. Chen, S. Xie, et al. “Masked Autoencoders Are Scalable Vision Learners”. In: (Nov. 2021).

[84] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: (Oct. 2020).

[85] Y. Li, D. Shamonin, T. Hassanzadeh, et al. “OP0002 EXPLORING THE USE OF ARTIFICIAL INTELLIGENCE IN PREDICTING RHEUMATOID ARTHRITIS, BASED ON EXTREMITY MR SCANS IN EARLY ARTHRITIS AND CLINICALLY SUSPECT ARTHRALGIA PATIENTS”. In: *Annals of the Rheumatic Diseases* 82.Suppl 1 (2023), pages 1–2. eprint: https://ard.bmj.com/content/82/Suppl_1/1.1.full.pdf.

[86] M. Firouzi, S. Fadaei, and A. Rashno. “A New Framework for Canny Edge Detector in Hexagonal Lattice”. In: *International Journal of Engineering* 35.8 (2022), pages 1588–1598. eprint: https://www.ije.ir/article_148876_abab35da30c6fdcd43e25b82efe9204d.pdf.

[87] Z. Zhou, V. Sodha, J. Pang, et al. “Models Genesis”. In: (Apr. 2020).

[88] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A Simple Framework for Contrastive Learning of Visual Representations”. In: Feb. 2020.

[89] A. van den Oord, Y. Li, and O. Vinyals. “Representation Learning with Contrastive Predictive Coding”. In: (July 2018).

[90] F. Schroff, D. Kalenichenko, and J. Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: (Mar. 2015).

[91] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pages 839–847.

[92] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia. “Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer”. In: *Procedia Computer Science* 179 (2021). 5th International Conference on Computer Science and Computational Intelligence 2020, pages 423–431.

[93] S. Maqsood, R. Damasevicius, and F. M. Shah. “An efficient approach for the detection of brain tumor using fuzzy logic and U-NET CNN classification”. In: *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part V 21*. Springer. 2021, pages 105–118.

[94] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929.

[95] X. Zhao, J.-W. Bai, Q. Guo, et al. “Clinical applications of deep learning in breast MRI”. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1878.2 (2023), page 188864.

[96] S. Shojaei, M. Saniee Abadeh, and Z. Momeni. “An evolutionary explainable deep learning approach for Alzheimer’s MRI classification”. In: *Expert Systems with Applications* 220 (2023), page 119709.

[97] M. Hussain, D. Koundal, and J. Manhas. “Deep learning-based diagnosis of disc degenerative diseases using MRI: A comprehensive review”. In: *Computers and Electrical Engineering* 105 (2023), page 108524.

[98] J. Hu, L. Shen, and G. Sun. “Squeeze-and-Excitation Networks”. In: *CoRR* abs/1709.01507 (2017). arXiv: 1709.01507.

[99] L. Folle, S. Bayat, A. Kleyer, et al. “Advanced neural networks for classification of MRI in psoriatic arthritis, seronegative, and seropositive rheumatoid arthritis”. In: *Rheumatology* (Mar. 2022).

[100] A. G. Howard, M. Zhu, B. Chen, et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *CoRR* abs/1704.04861 (2017). arXiv: 1704.04861.

[101] S. Mehta and M. Rastegari. “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer”. In: *CoRR* abs/2110.02178 (2021). arXiv: 2110.02178.

[102] R. Hemalatha, V. Vijaybaskar, and T. Thamizhvani. “Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning”. In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 233.6 (2019), pages 657–667.

[103] G. P. Avramidis, M. P. Avramidou, and G. A. Papakostas. “Rheumatoid arthritis diagnosis: Deep learning vs. humane”. In: *Applied Sciences* 12.1 (2022), page 10.

[104] X.-H. Li, C. C. Cao, Y. Shi, et al. “A survey of data-driven and knowledge-aware explainable ai”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pages 29–49.

[105] L. Saba, M. Biswas, V. Kuppili, et al. “The present and future of deep learning in radiology”. In: *European journal of radiology* 114 (2019), pages 14–24.

[106] B. Cheatham, K. Javanmardian, and H. Samandari. “Confronting the risks of artificial intelligence”. In: *McKinsey Quarterly* 2.38 (2019), pages 1–9.

[107] A. Bécue, I. Praça, and J. Gama. “Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities”. In: *Artificial Intelligence Review* 54.5 (2021), pages 3849–3886.

[108] M. T. Ribeiro, S. Singh, and C. Guestrin. “Anchors: High-precision model-agnostic explanations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Volume 32. 1. 2018.

[109] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Edited by I. Guyon, U. V. Luxburg, S. Bengio, et al. Volume 30. Curran Associates, Inc., 2017.

[110] V. Petsiuk, A. Das, and K. Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *CoRR* abs/1806.07421 (2018). arXiv: 1806.07421.

[111] G. Plumb, D. Molitor, and A. S. Talwalkar. “Model Agnostic Supervised Local Explanations”. In: *Advances in Neural Information Processing Systems*. Edited by S. Bengio, H. Wallach, H. Larochelle, et al. Volume 31. Curran Associates, Inc., 2018.

[112] M. T. Ribeiro, S. Singh, and C. Guestrin. “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pages 1135–1144.

[113] I. Ahern, A. Noack, L. Guzman-Nateras, et al. “NormLime: A New Feature Importance Metric for Explaining Deep Neural Networks”. In: *CoRR* abs/1909.04200 (2019). arXiv: 1909.04200.

[114] I. van der Linden, H. Haned, and E. Kanoulas. “Global Aggregations of Local Explanations for Black Box models”. In: *CoRR* abs/1907.03039 (2019). arXiv: 1907.03039.

[115] G. Montavon, S. Lapuschkin, A. Binder, et al. “Explaining nonlinear classification decisions with deep Taylor decomposition”. In: *Pattern Recognition* 65 (2017), pages 211–222.

[116] M. D. Zeiler and R. Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pages 818–833.

[117] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806* (2014).

[118] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR. 2017, pages 3319–3328.

[119] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[120] A. Shrikumar, P. Greenside, and A. Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMLR. 2017, pages 3145–3153.

[121] P.-T. Jiang, C.-B. Zhang, Q. Hou, et al. “Layercam: Exploring hierarchical class activation maps for localization”. In: *IEEE Transactions on Image Processing* 30 (2021), pages 5875–5888.

[122] R. Fu, Q. Hu, X. Dong, et al. “Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs”. In: (Aug. 2020).

[123] H. Wang, Z. Wang, M. Du, et al. “Score-CAM: Score-weighted visual explanations for convolutional neural networks”. In: (2020), pages 24–25.

[124] D. Omeiza, S. Speakman, C. Cintas, and K. Weldemariam. “Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models”. In: *CoRR* abs/1908.01224 (2019). arXiv: 1908.01224.

[125] R. Naidu and J. Michael. “SS-CAM: Smoothed Score-CAM for Sharper Visual Feature Localization”. In: *CoRR* abs/2006.14255 (2020). arXiv: 2006.14255.

[126] R. Naidu, A. Ghosh, Y. Maurya, et al. “IS-CAM: Integrated Score-CAM for axiomatic-based explanations”. In: *CoRR* abs/2010.03023 (2020). arXiv: 2010.03023.

[127] M. B. Muhammad and M. Yeasin. “Eigen-cam: Class activation map using principal components”. In: *2020 international joint conference on neural networks (IJCNN)*. IEEE. 2020, pages 1–7.

[128] Y. Zeng, J. Peng, X. Wu, and J. Hu. “Multi-CAM: A Class Activation Mapping Method Based on Multi-scale Feature Fusion”. In: *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE. 2022, pages 294–298.

[129] C. Zheng and W. Li. “Fusion-CAM: Fine-Grained and High-Faithfulness Explanations for Deep Convolutional Network via Hierarchical Fusion”. In: *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Volume 7. IEEE. 2024, pages 445–450.

[130] J. Gildenblat. *PyTorch library for CAM methods*. https://github.com/jacobgil/pytorch-grad-cam. 2021.

[131] D. Zhang, J. Han, G. Cheng, and M.-H. Yang. “Weakly supervised object localization and detection: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pages 5866–5885.

[132] F. Shao, L. Chen, J. Shao, et al. “Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey”. In: *Neurocomputing* 496 (2022), pages 192–207.

[133] Y. Yang and Y. Li. “A Lightweight Model for Physiological Signals-based Sleep Staging with Multi-Class CAM for Model Explainability”. In: *IEEE Sensors Journal* (2024).

[134] J. Adebayo, J. Gilmer, M. Muelly, et al. “Sanity checks for saliency maps”. In: *Advances in neural information processing systems* 31 (2018).

[135] O. Russakovsky, J. Deng, H. Su, et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pages 211–252.

[136] Microsoft. *Cats-vs-Dogs*. `www.kaggle.com/datasets/shaunthesheep/microsoft-catsvsdogs-dataset`.

[137] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pages 2278–2324.

[138] W. Stomp, A. Krabben, D. van der Heijde, et al. “Aiming for a shorter rheumatoid arthritis MRI protocol: can contrast-enhanced MRI replace T2 for the detection of bone marrow oedema?” In: *European radiology* 24 (2014), pages 2614–2622.

[139] S. G. Armato III, G. McLennan, L. Bidaut, et al. “The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans”. In: *Medical physics* 38.2 (2011), pages 915–931.

[140] X. Wang, Y. Peng, L. Lu, et al. “ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[141] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. “Dataset of breast ultrasound images”. In: *Data in Brief* 28 (2020), page 104863.

[142] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, et al. “A patient-centric dataset of images and metadata for identifying melanomas using clinical context”. In: *Scientific data* 8.1 (2021), page 34.

[143] I. Cherepanov, D. Sessler, A. Ulmer, et al. “Towards the Visualization of Aggregated Class Activation Maps to Analyse the Global Contribution of Class Features”. In: *World Conference on Explainable Artificial Intelligence*. Springer. 2023, pages 3–23.

[144] H. Park, J. Yun, S. M. Lee, et al. “Deep learning–based approach to predict pulmonary function at chest CT”. In: *Radiology* 307.2 (2023), e221488.

[145] F. P. Kroon, P. G. Conaghan, V. Foltz, et al. “Development and reliability of the OMERACT thumb base osteoarthritis magnetic resonance imaging scoring system”. In: *The Journal of rheumatology* 44.11 (2017), pages 1694–1698.

[146] N. Heller, F. Isensee, D. Trofimova, et al. *The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT*. 2023. arXiv: 2307.01984 [cs.CV].

[147] E. C. Newsum, A. H. van der Helm-van Mil, and A. A. Kaptein. "Views on clinically suspect arthralgia: a focus group study". In: *Clinical Rheumatology* 35 (2016), pages 1347–1352.

[148] D. Yuan, E. Rastogi, G. Naik, et al. "A continued pretrained llm approach for automatic medical note generation". In: *arXiv preprint arXiv:2403.09057* (2024).

[149] S. Goyal, E. Rastogi, S. P. Rajagopal, et al. "Healai: A healthcare llm for effective medical documentation". In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 2024, pages 1167–1168.

[150] G. Samarasinghe, M. Jameson, S. Vinod, et al. "Deep learning for segmentation in radiation therapy planning: a review". In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pages 578–595.

[151] A. C. Erdur, D. Rusche, D. Scholz, et al. "Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives". In: *Strahlentherapie und Onkologie* 201.3 (2025), pages 236–254.

[152] D. Hao, Q. Li, Q.-X. Feng, et al. "SurvivalCNN: A deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables". In: *Artificial Intelligence in Medicine* 134 (2022), page 102424.

[153] R. Aggarwal, V. Sounderajah, G. Martin, et al. "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis". In: *NPJ digital medicine* 4.1 (2021), page 65.

[154] X. Jiang, Z. Hu, S. Wang, and Y. Zhang. "Deep learning for medical image-based cancer diagnosis". In: *Cancers* 15.14 (2023), page 3608.

[155] P. Gamble, R. Jaroensri, H. Wang, et al. "Determining breast cancer biomarker status and associated morphological features using deep learning". In: *Communications medicine* 1.1 (2021), page 14.

[156] A. Echle, N. T. Rindtorff, T. J. Brinker, et al. "Deep learning in cancer pathology: a new generation of clinical biomarkers". In: *British journal of cancer* 124.4 (2021), pages 686–696.

[157] V. Andrearczyk, V. Oreiller, M. Jreige, et al. "Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT". In: *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*. Springer. 2021, pages 1–21.

[158] G. Samarasinghe, M. Jameson, S. Vinod, et al. "Deep learning for segmentation in radiation therapy planning: a review". In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pages 578–595.

[159] K. A. Tran, O. Kondrashova, A. Bradley, et al. “Deep learning in cancer diagnosis, prognosis and treatment selection”. In: *Genome medicine* 13 (2021), pages 1–17.

[160] T. Liu, E. Siegel, and D. Shen. “Deep learning and medical image analysis for COVID-19 diagnosis and prediction”. In: *Annual review of biomedical engineering* 24.1 (2022), pages 179–201.

[161] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. “A review on deep learning in medical image analysis”. In: *International Journal of Multimedia Information Retrieval* 11.1 (2022), pages 19–38.

[162] K. Wang, S. Yin, Y. Wang, and S. Li. “Explainable deep learning for medical image segmentation with learnable class activation mapping”. In: *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*. 2023, pages 210–215.

[163] A. Singh, S. Sengupta, and V. Lakshminarayanan. “Explainable deep learning models in medical image analysis”. In: *Journal of imaging* 6.6 (2020), page 52.

[164] H. Kang, H.-m. Park, Y. Ahn, et al. “Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis”. In: *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. Volume 11599. SPIE. 2021, pages 119–131.

[165] D. C. Cirean, A. Giusti, L. M. Gambardella, and J. Schmidhuber. “Mitosis detection in breast cancer histology images with deep neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer. 2013, pages 411–418.

[166] A. Prasoon, K. Petersen, C. Igel, et al. “Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pages 246–253.

[167] H.-C. Shin, H. R. Roth, M. Gao, et al. “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning”. In: *IEEE transactions on medical imaging* 35.5 (2016), pages 1285–1298.

[168] H. R. Roth, L. Lu, A. Seff, et al. “A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17*. Springer. 2014, pages 520–527.

[169] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. “Brain tumor segmentation using convolutional neural networks in MRI images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pages 1240–1251.

[170] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. “Unet++: A nested u-net architecture for medical image segmentation”. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*. Springer. 2018, pages 3–11.

[171] J. Chen, Y. Lu, Q. Yu, et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[172] S. Iqbal, T. M. Khan, S. S. Naqvi, et al. “TBConvL-Net: A hybrid deep learning architecture for robust medical image segmentation”. In: *Pattern Recognition* 158 (2025), page 111028.

[173] J. Wang, K. Wang, Y. Yu, et al. “Self-improving generative foundation model for synthetic medical image generation and clinical applications”. In: *Nature Medicine* 31.2 (2025), pages 609–617.

[174] F. Santini, J. Wasserthal, A. Agosti, et al. “Deep Anatomical Federated Network (Dafne): An Open Client-server Framework for the Continuous, Collaborative Improvement of Deep Learning–based Medical Image Segmentation”. In: *Radiology: Artificial Intelligence* (2025), e240097.

[175] J. Kamiri, G. M. Wambugu, and A. M. Oirere. “Multi-task Deep Learning in Medical Image Processing: A Systematic Review”. In: (2025).

[176] Y. Sun, L. Wang, G. Li, et al. “A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks”. In: *Nature Biomedical Engineering* 9.4 (2025), pages 521–538.

[177] Y. Zheng, W. Gan, Z. Chen, et al. “Large language models for medicine: a survey”. In: *International Journal of Machine Learning and Cybernetics* 16.2 (2025), pages 1015–1040.

[178] J. Huang, L. Yang, F. Wang, et al. “Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with Monte Carlo arbitrary-masked mamba”. In: *Medical Image Analysis* 99 (2025), page 103334.

[179] L. Han, T. Tan, T. Zhang, et al. “Synthesis-based imaging-differentiation representation learning for multi-sequence 3D/4D MRI”. In: *Medical Image Analysis* 92 (2024), page 103044.

[180] I. Rafegas and M. Vanrell. “Color encoding in biologically-inspired convolutional neural networks”. In: *Vision research* 151 (2018), pages 7–17.

[181] J. Wang, S. Wang, and Y. Zhang. “Deep learning on medical image analysis”. In: *CAAI Transactions on Intelligence Technology* 10.1 (2025), pages 1–35.

[182] A. Parashar, R. Rishi, A. Parashar, and I. Rida. “Medical imaging in rheumatoid arthritis: A review on deep learning approach”. In: *Open Life Sciences* 18.1 (2023), page 20220611.

[183] W. Khan, S. Leem, K. B. See, et al. “A comprehensive survey of foundation models in medicine”. In: *IEEE Reviews in Biomedical Engineering* (2025).

[184] J. Ma, Y. He, F. Li, et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (2024), page 654.

[185] Y. Huang, X. Yang, L. Liu, et al. “Segment anything model for medical images?” In: *Medical Image Analysis* 92 (2024), page 103061.

[186] R. Godasu, D. Zeng, and K. Sutrave. “Transfer learning in medical image classification: Challenges and opportunities”. In: *Transfer* 5 (2020), pages 28–2020.

[187] J. Wu, Z. Wang, M. Hong, et al. “Medical sam adapter: Adapting segment anything model for medical image segmentation”. In: *Medical image analysis* 102 (2025), page 103547.

[188] Y. Zhang, T. Zhou, S. Wang, et al. “Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pages 129–139.

[189] A. Karthikesalingam and P. Natarajan. “AMIE: A research AI system for diagnostic medical reasoning and conversations”. In: *Google Research* (2024).

[190] J. Bajwa, U. Munir, A. Nori, and B. Williams. “Artificial intelligence in healthcare: transforming the practice of medicine”. In: *Future healthcare journal* 8.2 (2021), e188–e194.

[191] R. Najjar. “Redefining radiology: a review of artificial intelligence integration in medical imaging”. In: *Diagnostics* 13.17 (2023), page 2760.

[192] M. Chen, B. Zhang, Z. Cai, et al. “Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey”. In: *Frontiers in medicine* 9 (2022), page 990604.

List of publications

Journal articles during Ph.D phase

Y Li, T Hassanzadeh, DP Shamonin, M Reijnierse, AHM van der Helm-van, BC Stoel. "Rheumatoid arthritis classification and prediction by consistency-based deep learning using extremity MRI scans." in *Biomedical Signal Processing and Control*, 2024, 91: 105990.

Yanli Li, Denis P. Shamonin, Tahereh Hassanzadeh, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "Feature analysis for proper intensity scaling and feature distinction in class activation maps." (*under revision*)

Yanli Li, Xikai Tang, Denis P. Shamonin, Hessam Sokooti, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Johan H.C. Reiber, Berend C. Stoel. "Aggregation of Class Activation Maps for Explaining Deep Learning at a Population Level." (*submitted*)

Yanli Li, Dennis A. Ton, Denis P. Shamonin, Monique Reijnierse, Annette H.M. van der Helm-van Mil and Berend C. Stoel. "Automatic joint inflammation estimation from hand and forefoot MRI based on regression neural networks." (*submitted*)

Yanli Li, Denis P. Shamonin, Monique Reijnierse, Annette H.M. van der Helm-van Mil and Berend C. Stoel. "Auxiliary-branch CAM in deep learning models serves as a tool for discovering undefined image patterns of rheumatoid arthritis." (*in preparation*)

Y. Yang and **Y. Li**, "A Lightweight Model for Physiological Signals-Based Sleep Staging With Multiclass CAM for Model Explainability," in IEEE Sensors Journal, vol. 24, no. 17, pp. 27815-27823, 1 Sept.1, 2024, doi: 10.1109/JSEN.2024.3430009.

Y. Yang and **Y. Li**. "Deep learning models as learners for EEG-based functional brain networks," in Journal of Neural Engineering, 2025, 22(2): 026005.

Tahereh Hassanzadeh, Denis P Shamonin, **Yanli Li**, Doortje I Krijbolder, Monique Reijnierse, Annette HM van der Helm-van, Berend C Stoel. "A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis" in Biomedical Signal Processing and Control 2024, 88: 105612.

International conference proceedings

Yanli Li, Denis Shamonin, Tahereh Hassanzadeh, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "OP0002 exploring the use of artificial intelligence in predicting rheumatoid arthritis, based on extremity MR scans in early

arthritis and clinically suspect arthralgia patients." in Annals of the Rheumatic Diseases, 2023, 82: 1-2.

Yanli Li, Denis Shamonin, Tahereh Hassanzadeh, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "POS1376 AUTOMATIC SEGMENTATION-FREE INFLAMMATION SCORING IN RHEUMATOID ARTHRITIS FROM MRI USING DEEP LEARNING." in Annals of the Rheumatic Diseases, 2024, 83: 730-731.

Yanli Li, Denis Shamonin, Tahereh Hassanzadeh, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "A simple but effective training process for the few-shot prediction task of early rheumatoid arthritis from MRI." in Medical Imaging with Deep Learning, 2022.

Denis Shamonin, **Yanli Li**, Tahereh Hassanzadeh, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "OP0190 QUANTIFICATION OF TENOSYNOVITIS, SYNOVITIS AND BONE MARROW EDEMA IN RHEUMATOID ARTHRITIS FROM WRIST MRIs, BASED ON DEEP LEARNING." in Annals of the Rheumatic Diseases, 2024, 83: 134-135.

Tahereh Hassanzadeh, **Yanli Li**, Denis Shamonin, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "OP0182 PREDICTION OF TREATMENT RESPONSE IN PATIENTS WITH ARTHRALGIA AT RISK FOR DEVELOPMENT OF RA, BY DEEP LEARNING FROM MRI SCANS OF THE WRIST, HAND AND FOOT." in Annals of the Rheumatic Diseases, 2024, 83: 133-134.

Denis Shamonin, Tahereh Hassanzadeh, **Yanli Li**, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "POS0920 QUANTIFICATION OF TENOSYNOVITIS IN RA FROM WRIST MRIs, BASED ON DEEP LEARNING." in Annals of the Rheumatic Diseases 2023, 82, 770-771.

Tahereh Hassanzadeh, Denis Shamonin, **Yanli Li**, Dennis A. Ton, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel. "POS0154 A DEEP LEARNING MODEL TO LOCATE INFLAMMATORY CHANGES IN RHEUMATOID ARTHRITIS." in Annals of the Rheumatic Diseases 2023, 82, 298-299.

Acknowledgements

To be written.

Curriculum Vitae

Yanli was born in Chengdu, Sichuan, China on June, 1995. In 2013, he started his bachelor in the major of Electronic Engineering at Tianjing University and continue his master study from 2017 in the major of Control Science.

From 2021, he started his PhD study in the Division of Image Processing (Dutch abbreviation LKEB) under the Department of Radiology at Leiden University Medical Center (LUMC) in the Netherlands. His PhD research mainly focuses on RA diagnosis, prediction, explainable artificial intelligence, and potential use of deep learning in helping clincial research.