UNIVERSITÀ POLITECNICA DELLE MARCHE

Department of Information Engineering (DII)

Master of Science in Biomedical Engineering

Transfer Learning for Informative-Frame Selection in US Rheumatology Images

Supervisor: Prof. Emanuele Frontoni

Co-Supervisors: Sara Moccia, PhD

Author:

Irene Guidotti

1091064

Academic Year 2019 - 2020

A Gioia e Letizia

Forget your perfect offering
There is a crack in everything
That's how the light gets in.

Ringraziamenti

Termina così questo percorso universitario che, tra un ponte vacillante e l'altro, mi ricorda che punto di partenza ora c'è. Per arrivare qui è stato essenziale il supporto di qualcuno che desidero ringraziare.

Ringrazio, innzanzitutto, il Prof. Emanuele Frontoni, relatore di questa tesi per due motivi. Il primo, come vuole la tradizione, per avermi dato la possibilità di lavorare a questo progetto, il secondo perchè in questa "feccia" di giovani studenti lei non vede solo studenti, ma piccole promesse. Grazie.

Desidero esprimere la mia gratitudine a Sara e Maria Chiara che da dietro le quinte sono state delle perfette direttrici dei lavori. A Sara la mia più profonda ammirazione per le trascendenti capacità organizzative e dirigenziali. A Maria Chiara una grazie di cuore per la dedizione, l'entusiasmo, e la disponibilità di questi mesi. Siete state per me esempi di passione e perseveraza.

Un ringraziamento speciale alle colonne portanti di questo viaggio, mamma e babbo. Tra le poche cose che conosco è che siete il polo nord e il polo sud: se mi perdo so dove trovarvi.

Un grazie di cuore va a Sonia, Fabio e Paola, i miei punti di riferimento ai quali faccio appello nei momenti di disorientamento.
A Sonia per i saggi consigli dispensiati in questi anni e per aver risvegliato in me la curiosità di apprendere cose nuove. Grazie Sonia.
A Fabio, per ricordarmi in continuazione che "un tempo aveva una sorella". Grazie

perchè è così che torno alla vita conviviale nei miei momenti di isolamento prolungato.
In più un grazie per il tuo contributo materiale senza il quale non sarei qui a scrivere.
A Paola perchè in questi cinque anni il suo continuo e costante sostegno è stato per me fonte di energia. Sei la mia dose di adrenalina giornaliera.
Siete la mia famiglia e il dono più bello. Attraverso i vostri occhi mi sono vista più forte e intelligente che attraverso i miei, grazie.

Un rigraziamento particolare va a Valeria ed Elena, da sempre le mie compagne di viaggio. Grazie perchè la vostra presenza ha un valore inestimabile per il mio essere.

Voglio ringraziare Arianna, che nonostante i miei prolungati silenzi, mi ha sempre sostenuta, capita e incoraggiata. Sei il mio rifugio quando l'insicurezza si impadrona di me.

Ringrazio Letizia, Gioia, Giulia e Gloria i miei quattro punti cardinali. Anche se lontane, siete la mia seconda famiglia. A Gioia e Letizia, senza le quali tutto questo non sarebbe mai stato possibile. A Giulia per l'incredibile capacità che ha di rasserenare e a Gloria perchè la sua grinta è da sempre fonte di ispirazione. In questi ultimi due anni un pò sofferti, sapere di avervi nella mia vita mi ha aiutata più di quanto voi possiate mai immaginare.

Ringrazio Ivana, che con la sua immensa pazienza mi supporta e sopporta da una vita.

Un grazie particolare va ad Enea, Cocò e Paolo perchè nonostante i pochi ritrovi siete gli amici di una vita.

Un ringraziamento sincero lo rivolgo a Simone, Elisa, Strozza, Alice e Beatrice per essere stati il mio team preferito in questi ultimi due anni di università.

Un grazie speciale va ai miei nonni che mi guidano dall'alto, per essere il faro che guida i miei passi, anche nei momenti più bui.

Un sincero ringraziamento va al gruppo 'La domenica della Peppa' perchè con il vostro affetto, entusiasmo ed allegria avete riempito ogni domenica passata insieme, rendendo questo percorso più sereno. Un grazie di cuore va a nonna Peppa, mia compagna di stanza e di qualche nottata in bianco. Sei la roccia di questa famiglia.

Ringrazio zio Alvaro e zia Lilly per aver sempre creduto in me.

L'ultimo ringraziamento lo rivolgo a Davide, per essere stato dal primo fino all'ultimo giorno il mio porto sicuro. Ti ringrazio perchè credi in me più di quanto lo faccia io e per ricordarmi, ogni volta che perdo la bussola, quali sono le cose importanti. Grazie perchè sai ascoltarmi ogni volta che metto a nudo le mie paure, ma ancora di più perchè fai di me, ogni giorno, una persona migliore.

Grazie per essere stati parte integrante di questo viaggio.

Irene

Abstract

Rheumatoid arthritis is a chronic autoimmune disease that causes pain, swelling and stiffness in the joints. It can affect any joint, but the wrist and hand joints are affected early in the disease process.

An effective treatment of the disease requires an early diagnosis. In this scenario, joint ultrasound scanning has played a leading role in the last decade, not only in the early diagnosis of disease, but also in monitoring its progress. The main limitation is represented by the fact that ultrasonography is an operator-dependent imaging technique. During an ultrasound examination, the doctor's manual ability and clinical experience are the principal elements to acquire correct information; elements that a young clinician might not possess. In fact, young ultrasonography residents undergo to a long period of training before acquiring a complete ultrasound competency.

During a metacarpal head ultrasound scanning, the ultrasound scan is considered as informative if the bony interfaces of the hyaline cartilage (the chondrosynovial and osteochondral interfaces) are present as continuous and sharp edges. If the two interfaces are lacking a metacarpal head ultrasound frame is considered as not informative.

The selection of informative ultrasound scans is a time consuming procedure that requires attention and prone to human error, especially if the ultrasound exam is performed by young and inexperienced clinicians. In order to provide medical personnel with useful tools to cope with this problem, an interesting solution is the development of an automatic informative frame selection system.

Several approaches have been proposed in the literature but none of these have achieved such performance as to translate the algorithms into the actual clinical practice.

One possible solution is to develop a deep learning method that, as observed in the literature, it is valid and promising tool, that outperforms classic machine learning approaches. Despite the high performance that deep learning models offer, they remain "black boxes". No justifications are provided on the decision path that artificial networks have made before reaching the final conclusion.

Understood the advantages offered by the deep learning and given the need to have explainable clinical decisions, this work aims to build a more transparent and understandable algorithm capable of automatically selecting informative ultrasound scans through the use of convolutional neural networks.

The proposed method is based on transfer learning techniques which involves the use of a pre-trained network, and the possibility of adapting and transferring the weights in order to be able to use its knowledge to pursue new tasks. Six different transfer learning approaches were performed on three different convolutional neural networks pre-trained on a dataset of natural images: VGG16, Inception V3 and ResNet50. The images were classified into two different classes: informative and not informative.

The approach was validated on 10 models obtained through a leave 4-subjects out cross validation. The method has been shown to be robust with mean values of sensitivity and specificity of 0.99 for informative and not informative classes and for the convolutional neural network with the best performances. The VGG16 completely fine-tuned has a Receiver Operating Characteristics that encloses an area equal to 99 %. These results beat not only the approaches found in the literature but also the results obtained with a training from scratch of the VGG16. For the "black box" problem, the algorithm offers, through the use of the gradient-weighted Class Activation Mapping, a visual explanation of the choices made by the convolutional networks, making the convolutional models more interpretable and understandable.

The promising results combined with the transparency of the proposed method helps in reducing the manual selection of informative images and speeding up the training process to which young clinicians undergo.

Sommario

L'artrite reumatoide è una malattia cronica autoimmune che causa dolore, tumefazione e ringonfiamento nelle articolazioni. Può colpire qualsiasi articolazione, ma nei primi due anni di sviluppo le articolazioni prinicipalmente coinvolte sono quella del polso e quelle della mano come le articolazioni interfalangee.

Un trattamento efficace della malattia richiede una diagnosi precoce. In questo scenario l'ecografia articolare ha avuto nell'ultimo decennio un ruolo leader, non soltanto nel diagnosticare precocemente la malattia, ma anche nel monitorare l'andamento di quest'ultima. Il limite principale è rappresentato dal fatto che è uno strumento operatore-dipendente.Durante un'ecografia articolare si richiedono particolari doti di manualità ed esperienza clinica; elementi che potrebbe non possedere un giovane clinico. L'ultrasonografia richiede da parte dei medici un lungo periodo di training per acquisire una completa autonomia operativa.

Considerando un'ecografia articolare eseguita sulle teste metacarpali, un'immagine viene considerata informativa se la cartilagine ialina risulta delimitata dalle due interfaccie osee (interfaccia condrosinoviale e quella osteocondrale), in caso contrario non informative. La selezione di ecografie informative è una procedura che richiede tempo, attenzione e soggetta all'errore umano, specialmente se ad effettuarla sono giovani e inesperti clinici. Al fine di fornire al personale medico strumenti utili per far fronte a queste problematiche, una soluzione interessante è lo sviluppo di una strategia in grado di selezionare in modo automatico le ecografie informative.

In letteratura diversi approcci sono stati proposti ma nessuno di questi ha raggiunto performance elevate, tali da introdurre gli algoritmi nella pratica clinica.

Una possibile soluzione risiede nei metodi di deep learning che come osservato nella letteratura, si sono dimostrati validi, promettenti e migliori rispetto ai classici metodi di apprendimento automatico. Nonostante le elevate prestazioni che i modelli di deep learning offrono, rimangono comunque "scatole nere": non vengono fornite giustificazioni sul percorso decisionale che la reti artificiali hanno compiuto prima di arrivare alla definitiva conclusione.

Compresi i vantaggi offerti dal deep learning e data l'esigenza di avere decisioni cliniche trasparenti questo lavoro ha come obbiettivo la costruzione di un algoritmo capace di selezionare automaticamente ecografie informative attraverso l'utilizzo di reti convoluzionali, cercando di rendendere comprensibile e spiegabile l'esito finale della valutazione automatica dell'immagine.

Il metodo proposto si basa sulla tecninca del transfer learning che prevede l'utilizzo di una rete pre-addestrata, ovvero nella possibilità di adattare e transferire i pesi al fine di poter utilizzare la sua conoscenza per perseguire nuovi obbiettivi. Sei diversi transfer learning sono stati effettuati su tre diverse reti convoluzionali pre-allenate su un dataset di immagini naturali: VGG16, Inception V3 e ResNet50. Le immagini sono state classificate in due diverse classi: informative e non informative.

L'approccio è stato validato su 10 modelli ottenuti attraverso un leave 4-subjects out cross validation. Il metodo si è dimostrato essere robusto ottenendo valori medi di sensitività e specificità del 0.99 per entrambi le classi e per la rete convoluzionale che ha ottenuto le migliori prestazioni. Utilizzando come classificatore la VGG16 completamente fine-tunata la Receiver Operating Characteristic racchiude un'area pari allo al 99 %. Questi risultati hanno superato non solo gli approcci presenti in letteratura ma anche i risultati ottenuti con un allenamento from scratch della rete VGG16. Per far fronte al problema della "scatola nera" l'algoritmo offre, attraverso l'utilizzo del gradient-weighted Class Activation Mapping, una spiegazione visiva sulle scelte effettuate dalle reti convoluzionali, rendendo i modelli convoluzionali più interpretabili e comprensibili.

I promettenti risultati uniti alla transparenza delle scelte effettuate dal metodo proposto potrebbero contribuire a diminuire l'intervento manuale richiesto per selezionare le immagini ed ad accellerare il processo di training a cui i giovani clinici vengono sottoposti.

Contents

Chapter 1

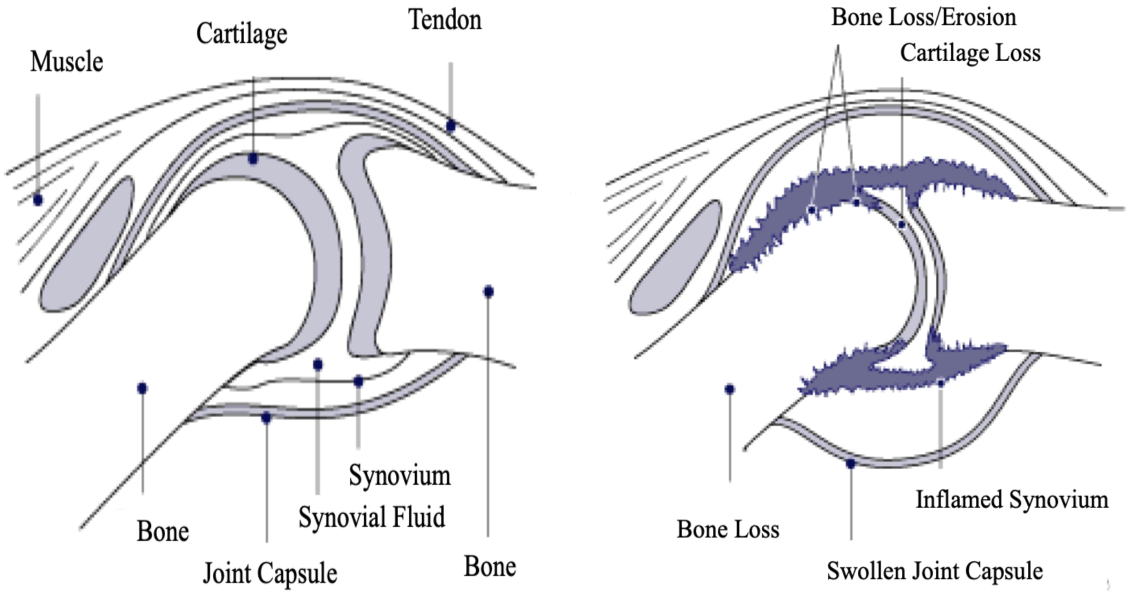
INTRODUCTION

Synopsis Rheumatoid arthritis is a chronic autoimmune disease that causes pain, swelling and stiffness in the joints. Early-stage diagnosis of rheumatoid arthritis is of primary importance to reduce and retard the progressive worsening of the disease. Ultrasound imaging is the most wide-spread technique for the management of rheumatoid arthritis. A series of advantages such as cost-effectiveness, non ionizing radiation, portability and, accessibility promoted the ultrasound diffusion and acceptance in rheumatology. However obtaining quality informative frames is a tedious task. Supporting the clinician during the informative frame selection is the most significant step for the successive diagnosis of the patient.

In this Chapter Section 1.1 introduces rheumatoid arthritis. The current imaging management of the disease and its limitation are discussed in section 1.2. Finally in section 1.3 the aim of the thesis is presented.

1.1 Rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic autoimmune disease that affects about 1% of the world population with a reported annual incidence of about 40 in 100000 worldwide [1]. An autoimmune disease is a condition in which the body's immune system mistakenly attacks the healthy tissues. In RA, the dysfunctional immune system primarily hits the lining tissue of the joints or synovium causing a chronic synovial inflammation, thickening of the synovial membrane, cartilage degradation, joint space narrowing, juxta-articular bone erosion, and extra articular manifestations (Figure 1.1).



(a) *Normal Joint.* **(b)** *Joint affected by RA.*

Figure 1.1: *(a) Structure of a normal and general joint. (b) Principal structural changes caused by RA in a general joint.*

The progressive spread of synovitis to the joint capsule and surrounding cartilage and bone causes deformity, loss of function, stiffness and pain in the joints; resulting in a disability and premature mortality [2]. Early RA typically manifests with signs of joint inflammation and is most likely to erode wrist, metacarpophalangeal, and interphalangeal joints [3, 4] . There are some reports of genetic heterogeneity between males and females in RA and the mean age of onset in females is significantly lower than males that suggests the presence of additional predisposing factor(s) in female patients [5]. Generally, the incidence of RA is two to three times higher in women than men. This observed sex-bias may be because of differences in biological factors, such as hormones or sex-related genetic factors [6].

The identification of RA at its initial stage is essential for preventing the progression of the disease [7]; in fact, an earlier treatment may change and retard the erosive effects of the disorder. The need of an early diagnosis brought in 1978 the American College of Rheumatology (ACR) to develop the first RA classification criteria; but it was criticized for lack of sensitivity in early disease.

In 2010, a joint working group of ACR and the European League Against Rheumatism (EULAR) developed a new approach to classify the RA [8]. In this work the factors that best discriminate subjects that are and are not at high risk for RA are defined. In this criteria set, classification as "definite RA" is based on:

1. The confirmed presence of synovitis in at least 1 joint
2. Absence of an alternative diagnosis that better explains the synovitis
3. Achievement of a total score of 6 or greater (of a possible 10) from the individual scores in 4 domains:
   (A) Number and site of involved joints (score range 0–5)
   (B) Serologic abnormality (score range 0–3)
   (C) Elevated acute-phase response (score range 0–1)
   (D) Symptom duration (score range 0–1)

The joint involvement of the domain (A) differs from the synovitis in 1 joint mentioned above. It refers instead to any joint presenting with active synovitis symptoms hence, swelling and or tederness. In domain (B), the serologic abnormality refer to an anomalous levels of the rheumatoid factor (RF) and of the anti-citrullinated protein antibody (ACPA). RF and ACPA are the two autoantibodies clinically useful for the RA diagnosis [9]. The C-reactive protein (CRP) and the erythrocyte sedimentation rate (ESR) quantify the acute-phase response of domain (C) and they are correlated with the severity of disease [10]. CRP and ESR are scored based on laboratory standards. The symptoms duration in domain (D) refers to the patient's self-report of the maximum duration of signs or symptoms of synovitis (pain, swelling, and tenderness) [8].

Since soft-tissues structural damages and synovitis occur earlier in the diseases [11], imaging techniques have acquired through the years a key role in the RA assessment and progression [12]. In a survey of 2012, which included 154 rheumatologists, the imaging examination was estimated as the most important element for RA evaluation [13]. The need of including imaging techniques in the early diagnosis of RA led the

EULAR organization to release, in 2013, 10 recommendations for the use of imaging in joints in the clinical management of RA [14]. Recommendations state that diagnostic certainty in RA is improved by imaging in comparison with the clinical examination, in addition to accurate assessment of joint inflammation, joint erosion, prediction of treatment response, and disease activity monitoring.

1.2 Diagnostic imaging

Conventional radiography (CR) has been considered the gold standard for imaging in RA, but its sensitivity for structural damage in early RA diagnosis is low [15]. CR is perfectly suitable in detecting bone erosion, joint space narrowing, and new bone formation, but they can not portrait inflammatory changes [16] and so disease activity cannot be assessed.

Over the last decade, the clinical management of RA has changed thanks to the improvement and advancement of ultrasonography [17].

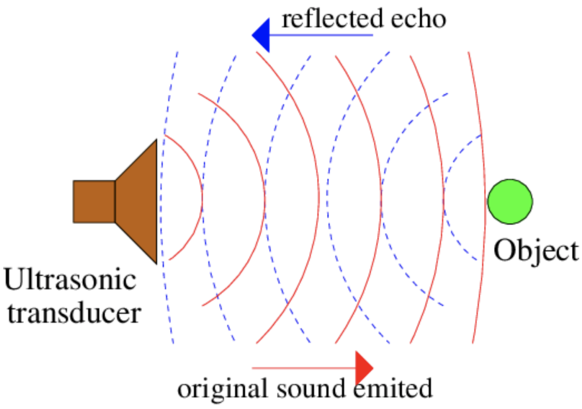


Figure 1.2: *Work principle of US imaging technique. The ultrasonic transducer acts as an emitter and receiver. The red lines represent the US beam produced by piezoelectrical crystals, while the dot blue lines are the echo waves reflected back to the crystals.*

Ultrasonography is a medical imaging device that uses high-frequency sound waves to view inside the body. The ultrasound (US) probe converts electrical energy to high-frequency sound via a transducer containing piezoelectrical crystals. These type of crystals can also work in reverse, they produce electrical signals when they detect high frequencies pressures sound waves (Figure 1.2). When the US probe directs the US beam into the body, it passes trough the skin into the internal anatomy. As the waves encounter tissues of different characteristics and densities, they produce echos that are reflected back to the piezoelectric crystals. The returning echoes are converted into electric signals at first and then, by a computer, into points of brightness on the image corresponding to the anatomic position and strength of the reflected echo. The US probe contains an array of crystals which allows to make a series of image lines that together form a complete image frame called sonogram. The ultrasound image is so produced based on the reflection of the waves off of the body structures.

Several studies [18, 19, 20, 21, 22, 23, 24], have underlined the potential and leading role that US has in the early detection of RA. According to EULAR imaging recommendations, the detection rate of synovitis at the hand and wrist using US was double than the that obtained with clinical evaluation [14]. Patients assessed by US are likely to fulfil the ACR/EULAR criteria for RA at an earlier stage of their disease than those assessed using convectional assessment [25, 26, 27].

The huge benefit offered by US is the excellent soft-tissue contrast that allows to depict RA at the very initial presentation. Furthermore, US offers a series of advantages over other imaging techniques such as CR, including the lack of ionizing radiation, non-invasiveness, portability, accessibility, and cost effectiveness.

The growing evidence of US benefits in the management of RA has encouraged many rehumatologists to embrace US in their clinical practice. In 2016, a group of US experts developed pragmatic suggestions for the use of US in the daily management of patients with suspected or established RA [28]. Despite its advantages, US presents also unique technical challenges. Indeed US is an operator-dependent imaging modality and before reaching US competency a significant learning process and period of training are required [29].

Considering the US scanning of the hyaline cartilage of a metacarpal head, which is one of the most affected by early RA [3, 4], the doctor manually moves the probe over the small joint (Figure 1.3). According to the definition of healthy hyaline cartilage of the Outcome Measure in Rheumatology US [30], a metacarpal head US frame is considered as informative (I) if the hyperechogenic osteochondral interface is detectable, even partial damaged and, if the hyperechogenic chondrosynovial interface

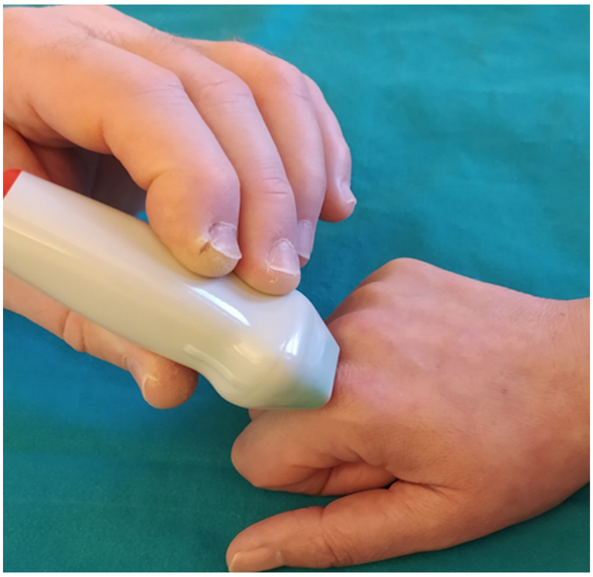
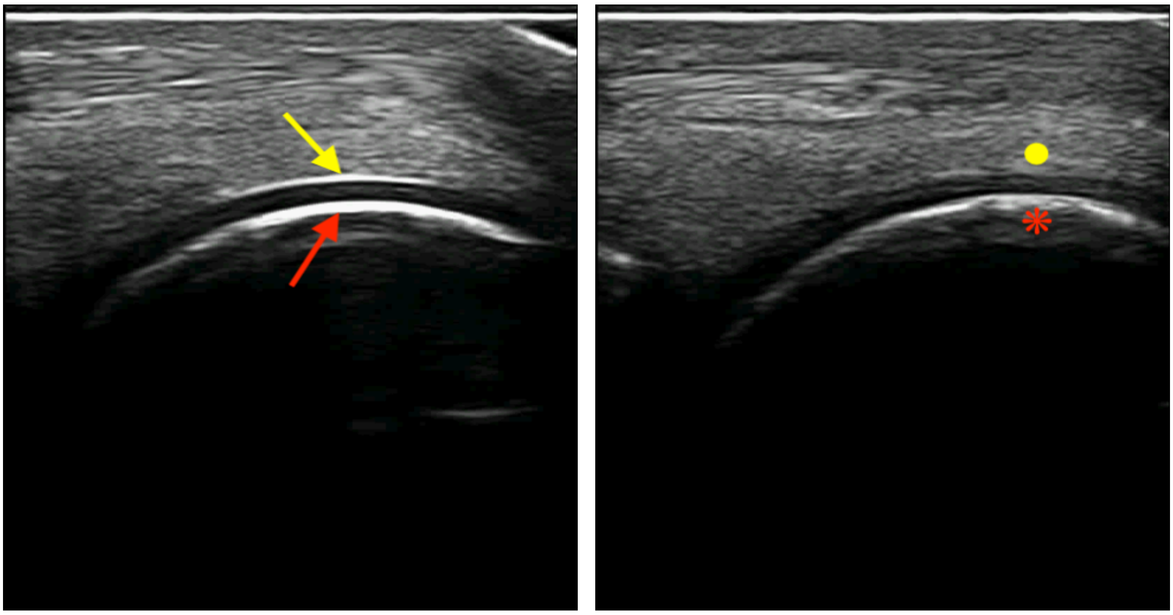


Figure 1.3: *The figure shows the positioning of the US probe during the metacarpal head scanning. The probe is moved along the skin surface of the small joint in a continuous contact.*

is present (it may not be visible in damage cartilage, though) (Figure 1.4a). When these two features are lacking a metacarpal head US frame is considered as not informative (NI) in healthy subjects. (Figure 1.4b).



(a) *Informative.*

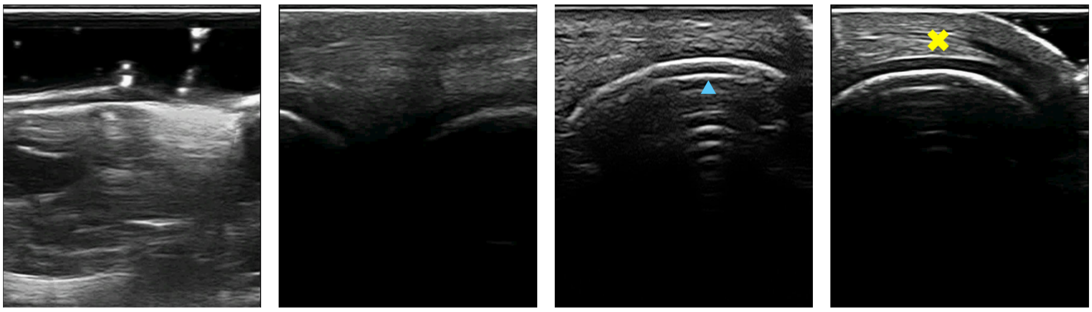
(b) *Not informative.*

Figure 1.4: *Samples of an I frame (a) and NI frame (b) of a metacarpal head joint. In (a) the hyaline cartilage appears as a homogeneous and delimited by two regular, sharp, and continuous hyperechoic interfaces. The yellow arrow represents the chondrosynovial interface while red arrow refers to osteochondral interface. In (b) the two interfaces are not visible.*

During the examination, the doctor's US scanning experience and manual ability are the principal elements to obtain and select quality I frames. However during the scanning proceeding, a series of drawbacks may potentially lead to the selection of unreliable US scans with a subsequent misdiagnosis. Multiple factors affect the correct performance and interpretation of I US frames, including:

- The correct scanning technique as the proper positioning of the transducer
- Machine settings, including the correct choice of US frequency
- Anatomy complexity (the overlapping of structures)
- Artifacts like acoustic shadow, reverberation and refraction.

These drawbacks contribute in challenging the collection of I frames by extending the time for the visual detection of abnormalities and/or artifacts and by complicating the correct discrimination of I and NI frames. Examples of NI frames are displayed in Figure 1.5. In particular, in frame Figure 1.5c is possible to observe how a reverberation artifact produces, just beneath the chondrosynovial interface, a white curved line indicated by a small blue triangle. At first glance and for an young US resident doctor this reverberation artifact can be mistaken as one of the two hyperechoic interfaces. In frame Figure 1.5d the chondrosynovial interface is partially visible and so the frame might be mystified as an I one.



(a) *NI frame* **(b)** *NI frame* **(c)** *NI frame* **(d)** *NI frame*

Figure 1.5: *The figure shows examples of NI frames of the metacarpal head joint. In (a) and (b) the metacarpal head does not appear. In (c) the chondrosynovial interface is not displayed and the small light blue triangle evidences the reverberation artifact. In (d) the chondrosynovial interface is only partially visible indicated by the yellow cross.*

The current solution for the selection I frames is carried out by the human visual system, resulting in a subjective, time consuming and qualitative operation. From the prospective of image analysis, it is essential to develop an advanced automatic US image frame selection that assists the clinician in the selection of I frames. An automatic selection of I frames can potentially improve the doctor's work, by reducing the time for the visual selection of quality images and by lowering the variability in interpretation of results for a more objective diagnosis. Moreover, it may help and speed up the young residents' training, in quality-control standardization and detection of the cartilage damage [31].

1.3 Computer-assisted US image analysis

To accomplish the frame selection task several artificial intelligence (AI) approaches such as machine learning (ML) and deep learning (DL), have been proposed. In Chapter 2 a details description of the state of art is provided.

Although AI has been around since the 1950s, only recently it has been introduced in medical imaging. AI techniques are an assortment of mathematical algorithms capable of identifying patterns in data and performing predictions on new information.

Automatic frame selection that employs ML or DL techniques involves supervised learning. In supervised learning a classifier is trained with labeled inputs to develop a predictive model. During the training process, these algorithms learn underlying patterns within the data and exploit them to predict unseen images.

ML algorithms work on human designed features, named handcrafted features (i.e. texture features, morphological features, frequency features); while DL models learn features by theirself during the training phase.

In the medical imaging context a variety of DL structures have been explored. Among them Convolutional Neural Networks (CNNs) are the most popular choices for image classification. A CNN consists of convolutional layers, capable of automatically extract features at pixel-level data.

With the development of DL, researchers noted that features extracted by the deep neural networks outperform features designed by humans [32, 33]. Indeed, in DL models the learned features adapt and change based on the input; while in ML the handcrafted features, once they have been defined, are independent from the input images. Although the ability of automatically learn representative features, CNNs require millions of labeled images before reaching satisfactory performances. In healthcare collecting such high number of annotated images is expensive, and moreover privacy regulations limit their usage [34]. To address this issues one of the most commonly used methods is transfer learning. Unlikely ML that learns each task from scratch [35], DL allows to adapt and exploit the knowledge of pre-trained architecture to solve new problems. Despite the high performances that DL models offer, they remain "black boxes". In fact, no justifications are provided on the decision path that artificial networks have made before reaching the final conclusion. This drawback has limited the actual introduction of DL models in the clinical practice.

1.4 Aim of the thesis

US has undergone to a dramatic evolution in the field of rheumatology, due to its reliability and efficiency in displaying the early joint effects caused by RA. However, a proper scanning proceeding requires years of experience, manual ability and an attentive visual inspection. Consequently selecting quality US frames results to be time consuming and prone to human errors.

Motivated in providing a reliable trusty support during the US examination, this thesis addresses the problem of automatic I frame selection in metacarpal head scanning. To overcome limitations of ML algorithms (Chapter 2), the proposed approach exploits the potentiality of DL. Furthemore, it deals with the problem of AI interpretability, through the use of the gradient-weighted Class Activation Mapping.

The thesis aims to develop a trusty DL system, by operating a transfer learning strategy with three different pre-trained CNNs. Different levels of transfer learning will be performed to investigate if the features learned by CNNs on natural images can be exploited for the I frame selection. Visual explanations of the choices made by the CNNs are provided in order to built a trusty and synergic system between the clinicians and the algorithm. A flowchart of the proposed approach is present in Figure 1.6

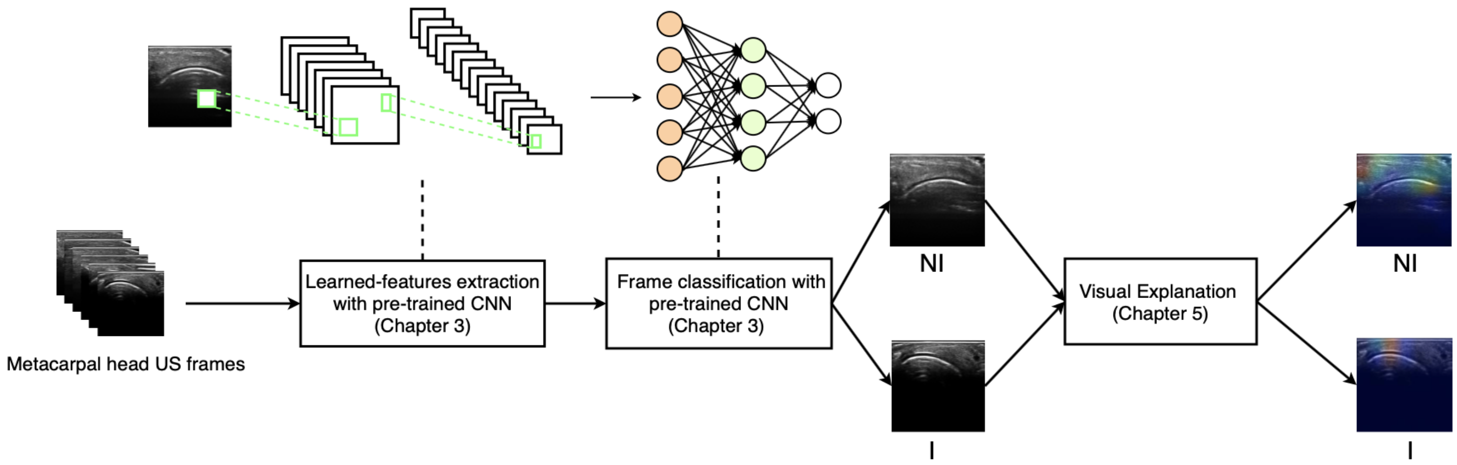

Metacarpal head US frames
Learned-features extraction with pre-trained CNN (Chapter 3)
Frame classification with pre-trained CNN (Chapter 3)
NI
I
Visual Explanation (Chapter 5)
NI
I

Figure 1.6: *Flowchart of the proposed approach to automatic frame selection in metacarpal head US scanning.*

1.4.1 Structure of the thesis

The thesis is structured as follow:

Chapter 2 summarizes the state-of-art approaches for informative frame selection pointing out the main limitations.

Chapter 3 describes the proposed approach for the I frame selection problem exploiting CNNs.

Chapter 4 deals with the experimental protocol used in the proposed methodology. In the chapter the data splits for the training, validation and testing set are described. The choices relative to the parameters tuning for the training procedure are meticulously reported together with technical specification about the used platforms.

Chapter 5 details the results obtained with the proposed method. The chapter has three subsection describing respectively the numerical transfer learning results of the proposed approach, the results of the CNN trained from scratch, and the graphical results of the gradient-weighted Class Activation Mapping.

Chapter 6 discusses the results obtained in Chapter 5 including the main limitations of the proposed method and the impact that the method can have in the actual clinical practice.

Chapter 7 concludes the thesis summarizing the proposed method and its strength. The chapter also introduces future developments that can be carried out.

Chapter 2

RELATED WORK

Synopsis The principal solutions for the automatic frame selection systems are reported and discussed in this Chapter. The proposed approaches can be divided in two, depending on the feature extraction process. Section 2.1.1 describes the threshold-based and ML methods built on handcrafted features, while DL applications are reported in Section 2.1.2. The main limitations of the state-of-art frame selection are discussed in Section 2.2.

2.1 Current solutions for automatic frame selection

Considering that medical imaging supplies important information of an organ function and anatomy, it is essential to collect useful frames in order to detect the state of a possible disease. During the medical images processing, the manual selection of informative frames is still the gold standard. In recent years though, the topic of frame selection has showed an increasing will of developing an automatic system. The benefits carried out by an automatic classification of frames of interest (frames with informative content) are:

- Increased repeatably
- Relieving doctor's workload
- Reducing futile image analysis

Although automatic frame selection in US joint scanning is relative unexplored, different solutions have been proposed for video summarization in endoscopy procedures.

In literature the strategies proposed for the automatic frame selection can be divided into two branches, based on the feature extraction modality: handcrafted features and learned features.

2.1.1 Automatic frame selection using handcrafted features

Since frame selection is a classification problem, the main challenge consists in identifying patterns in images that define the class of the frame. Handcrafted features are measurable properties of an image and they can focus on different characteristics as edges, corner, colors, texture. These features are manually designed and they try to attempt to model the features that the doctor look when identifying an useful frame.

ML and threshold sensitive approaches are the two solutions proposed in literature that address the problem of automatic frame selection using handcrafted features.

In [36] two techniques are investigated for the selection of useful frames during a colonoscopy. The first technique exploits the Canny edge detector (Figure 2.1), an operator that uses a multi-stage algorithm to detect a wide range of edges in images. Successively the images are classified based on a threshold method. The performance of this edge-based technique is sensitive to the selected threshold value.
In the second approach seven texture features (i.e., entropy, contrast, correlation, homogeneity, dissimilarity, angular second moment, and energy) derived from the gray level co-occurance matrix (GLCM) are classified with a k-means clustering.

In a more recent work, [37] proposed for the colonoscopy summarization, a two-step algorithm. In the first step, an edge threshold based approach was implemented to discriminate uninformative and ambiguous frames. In the second step, the ambiguous frames are classified in informative or not informative using information derived from the brightness segmentation feature extraction. The performances of this approach are susceptible to the threshold set in the edge detection step.

A random forest classifier was used by [38] for the same application. The features investigated for each frame were: corner and edge features matched with the previous frame, the percentage of edge pixels, and the mean and standard deviation of intensity

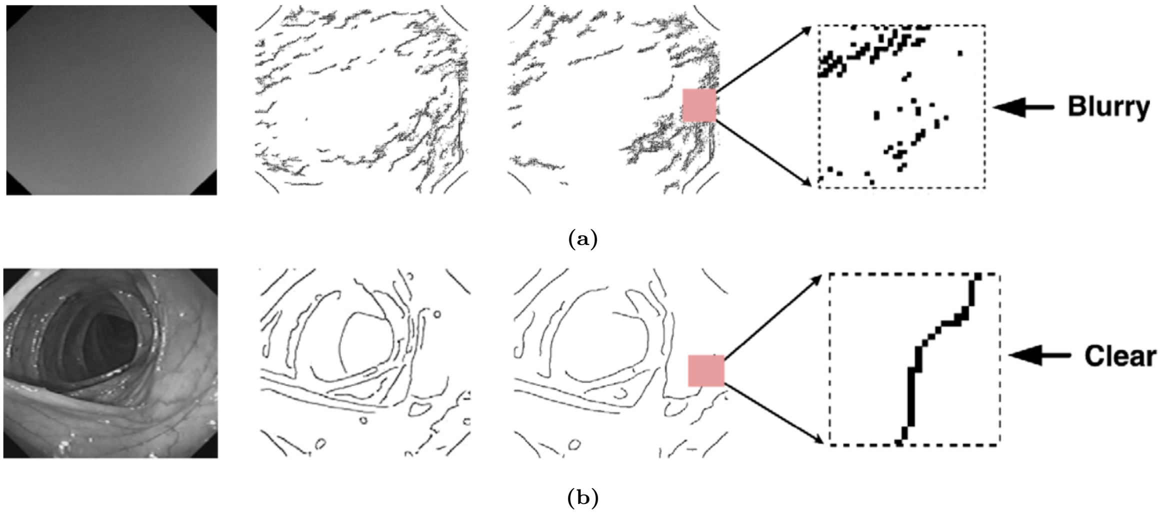


Figure 2.1: *Samples of edges detected by the Canny edge detector from (a) an uninformative frame and (b) from an informative frame. As shows the figure the edges generated by the Canny edge detector in the uninformative frame are blurry while those of the informative one are well defined. This figure is courtesy of [36]*

in hue-saturation-value (HSV) color space.

Wireless capsule endoscopy (WCE) is a novel non-invasive procedure that allows the visualization of the gastrointestinal tract as a mean to identify possible illness. The major drawback of this technology is the excessive amount uninformative frames for video diagnosis. Therefore, [39, 40, 41] proposed three different methods for the automatic detection of informative frames.

In [39] a ML approach was employed for the removal of uninformative frames. A support vector machine (SVM) was trained with local color features (moments and histogram) to discriminate between potentially informative frames and high bubble contaminated frames.

In [40] a threshold approach was used to select informative frames. For each frame was computed a saliency map obtained with three features: image moments (mean, standard deviation, skewness, and kurtosis), multi-scales contrast, and curvature.

In [41] the automatic classification is focused on informative and uninformative frame regions. First, the gray scale image is generated from the red-green-blue (RGB) frame, then morphological operations are designed to emphasize the round objects in a frame. The subsequent operations involve: a median filter for noise reduction, a fuzzy k-means clustering to find lighter regions, and a sigmoid function to select region of interest in which textures features are extracted. Finally a HSV threshold called HSV discriminator is applied to classify informative and not informative regions.

The work in [42] presents an automated frame selection algorithm for high-resolution microendoscopy video sequences. Images free of motion artifact and with sufficient intensity for meaningful analysis are selected according to their intensity, entropy and the number of keypoints (i.e., point of interest detected on the image). Then, results of the automated process were statistically compared to manual frame selection done by a trained observer.

The work of [43] employed a Gaussian mixture model for the automatic detection and removal of uninformative frames in pulmonary optical microendoscopy. In this approach texture features were extracted from the GLCM.

In [44], the problem of a robust and automatic classification of informative frames is focused on laryngoscopy and it exploits the power of ML. Intensity features, keypoints, and spatial content features were selected and used in a multi-class SVM. Frames were classified as informative, blurred, with saliva or specular reflections and underexposed.

2.1.2 Automatic frame selection using learned features

In this section, DL methods are advanced for the development of an automatic frame selection system. In DL models, the deep layers act as feature extractors. These features are directly learned from the input images, and they are named learned features.

In [45], CNNs are applied for the automatic selection of uninformative frames to be discarded in colonoscopy procedures with the purpose of video summarization. Two strategies were developed and compared. In the first strategy authors developed a CNN, named SimpleNet, trained from scratch. In the second strategy a fine-tuning approach was adopted and applied on CNNs pre-trained on ImageNet[1] (a dataset for natural-image classification, made of more of 14-million images). The exploited architectures were AlexNet, GoogleNet, and ResNet50. In this work ([45]), the fine tuning approach outperforms the learning from scratch as already demonstrated several times in the literature.

[1] `http://www.image-net.org`

The work in [46] presented an automatic selection of informative frames in laryngoscopic procedures. In this work two consecutive transfer learning approaches were adopted and compared with a learning from scratch. In the first approach several CNNs (VGG16, InceptionV4, ResNet V1 101 and ResNet V2 102, ResNet V2 152, and Incetpion-ResNet V2) were exploited as features extractor and the extracted learned features were then classified by means of SVMs. In the second transfer learning approach the best performing CNN (coming from the first approach) was fine-tuned and it was used both as feature extractor and classifier. For comparison purposes, the same best performing CNN was trained from scratch. In this work the VGG16 results to be the best CNN. In details, the fine tuned VGG16-SVM classification and VGG16-based classification achieved comparable performances both higher than VGG16 trained from scratch.

To the best of author's knowledge, the work in [31] is the first attempt in the field of US joint scanning to address the problem of automatic classification of I frames. The proposed approach exploited the potentiality of transfer learning for the VGG16 to classify I frames and NI frames. In this work the I and NI frames refer to the frames previously discussed in Section 1.2.

Table 2.1 summarizes the aforementioned state-of-the-art approaches to informative frame selection.

Table 2.1: *State-of-the-art approaches to informative-frame selection.*

Method	**Year**	**Anatomical District**	**Feature set**	**Classification**
Oh et al. [36]	2007	Colon	Texture and edge	k means
Bashar et al. [39]	2010	Gastro-intestinal tract	Intensity and texture	Support vector machines
Mehmood et al. [40]	2014	Gastro-intesinal tract	Image moment, curvature and color histogram	Threshold-based approach
Maghsoudi et al. [41]	2014	Gastro-intesinal tract	Intensity	Threshold-based approach
Ballesteros et al. [37]	2015	Colon	Intensity and edges	Threshold-based approach
Armin et al. [38]	2015	Colon	Motion, intensity and image derivatives	Random forest
Ishijima et al. [42]	2015	Esophagus	Intensity, entropy and key-points	Statistical comparison
Perperidis et al. [43]	2016	Lungs	Texture	Gaussian mixture model
Moccia et al. [44]	2018	Larynx	Intensity, entropy, key-points and texture	SVMs
Islam et al. [45]	2018	Colon	Learned features	CNN
Fiorentino et al. [31]	2019	Metacarpal head	Learned features	CNN
Patrini et al. [46]	2020	Larynx	Learned features	CNN and SVMs

2.2 Limits in the literature

The majority of the methods described in Section 2.1.1 do not achieve the expected performances for the actual application in the clinical practice. Indeed, the threshold-base approaches are susceptible to the threshold value, and they can not compete with high inter-subject variability typical of medical data.

Concerning the ML algorithms, they are trained and tested on small and local dataset and they may not necessarily perform as well with new data. Moreover, finding the informative, discriminating and independent set of features for the ML training is a complex task and it requires tremendous efforts. It is difficult to define a mathematical model for extracting features that are intuitive and immediate to humans.

In the last decade, it has been shown that DL algorithms outperform ML in image analysis [32, 33, 47]. The huge benefit of learned features is their automatic definition. During the training process they change and adapt based on the input image and the corresponding label, without being manually designed.

DL have showed promising results in the field of the medical imaging for breast lesion classification [48, 49] using US images, for skin cancer classification [50] with dermoscopy images, but also for the semantic segmentation of anatomical structures in magnetic resonance images [47].

In [45, 46, 31], the applied transfer learning method enable to reach encouraging results, despite the limited amount of data. In fact in DL, thanks to the transfer learning technique, the knowledge of a trained network can be reused for a new task in which the reduced amount of data does not permit a complete training of the CNN (Chapter 3). In ML the knowledge of a classifier can not be reused and each task is learned from scratch.

The main drawback of DL algorithms is the interpretability. CNNs make the final decision using high-levels features, difficult to be explained to humans. In order to build trust in intelligent systems, and to move towards their meaningful integration into the clinical practice, it is clear that there is the need of building 'transparent' models that have the ability to explain why they predict and what they predict. With the purpose of making CNNs more transparent and explainable [51] introduced in 2017, a method called gradient-weighted Class Activation Mapping (grad-CAM, Chapter 3) for the interpretation of the CNN final output.

Motivated by the need of an automatic frame selection system and by the advantages of DL use, this thesis is the first work that attempts to built a more transparent algorithm using interpretability techniques and exploring the efficiency of different transfer learning approaches for the metacarpal head US frame selection.

Chapter 3

METHODS

Synopsis This Chapter introduces and discusses our approach developed for the automatic I frame selection of US metacarpal heads. After a brief overview on CNN in Section 3.1, the Chapter proceeds illustrating the employed transfer learning strategy (Section 3.2.1) and the training settings (Section 3.2.2). For building a more trusty system while using CNNs, it has been resorted to the grad-CAM described in Section 3.2.4.

3.1 Overview on convolutional neural networks

CNNs are perfectly suitable for image classification and they have been employed in most of the developed automatic frame selection systems. CNNs are deep hierarchical neural models that roughly mimic the nature of mammalian visual cortex, and are the most promising architectures for such task [52]. The base of a CNN is the convolution which is a mathematical operation of two functions that produces a new function. The convolution theorem states that under certain conditions the Fourier transform of a convolution is a point-wise product of Fourier transforms. In other words, convolution in one domain (i.e., time domain) equals point-wise multiplication in the other domain (i.e., frequency domain). Considering a one dimensional application the convolution can be expressed as 3.1:

$$g(x) = f(x) \circledast h(x) = \int_{-\infty}^{\infty} f(s)h(x-s)ds \tag{3.1}$$

where *f(x)* and *h(x)* are two functions and s the dummy variable. In the image processing context f indicates the input (the image), h refers to the kernel (filter), and g identifies the feature map. The standard architecture of a CNN consists of alternating:

- Convolutional layers
- Activation functions
- Pooling or subsampling layers

and it ends with:

- Fully-connected layers

Their operating principle are hereby reported:

Convolutional layer The purpose of a convolutional layer is to extract features from the input layer. These layers are comprised of a series of filters or learnable kernels which aim at extracting local features from the input, and each kernel is used to calculate a feature map or kernel map. The first convolutional layer extracts low-level meaningful features such as edges, corners, textures and lines. Next convolutional layers extract higher-level features, but the highest-level features are extracted in the last convolutional layer. Each kernel is a matrix, spatially smaller than the image, which convolves around the feature map or the input image (Figure 3.1). Kernel size refers to the size of the kernel matrix. The kernels are not predefined but learned during the training.

During the convolutional procedure, three hyperparameters control the size of the output volume: the depth, the stride and the zero-padding. The depth corresponds to the number of the used filters. The higher the number of kernels of the layer, the higher the channel number of the output volume and the amount of extracted features. The stride is step size with which the kernel slides over the image, and it will produce

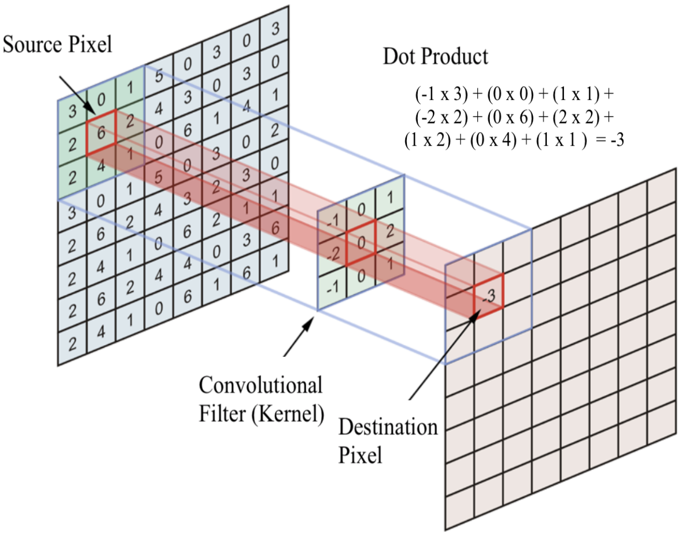

Source Pixel
Dot Product
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(1 x 2) + (0 x 4) + (1 x 1) = -3
-3
Convolutional
Filter (Kernel)
Destination
Pixel

Figure 3.1: *Example of an image convolution. The 3x3 kernel convolves around the 7x7 input image, computing at each step the dot product. The process is repeated for every pixel in the image. The source pixel is the anchor point at which the kernel is centered.*

smaller output volumes spatially. The zero-padding pads zeros pixels to the border of the generated feature map.

Activation functions Activation functions consist of non-linear layers that take the feature map generated by the the convolutional layer and creates an activation map. The activation function introduces the non-linearity into neural networks and it allows the learning of more complex features. There are several nonlinear activation functions such as *tanh(x)*, *sigmoid(x)*, and Rectified Linear Unit ($ReLU$).

Pooling (or Subsampling) layer The pooling or subsampling layer reduces the resolution of the feature maps through compressing features and computational complexity of the network. The most common pooling is the max pooling, but even the average pooling is often implemented. An example of max pooling and average pooling is shown in (Figure 3.2).

Multiple sequential convolution, kernel, and pooling steps result in numerous layers of data (pooled features maps) that are transformed into a 1-directional array through a process called flattening to be further process as input of dense layers.

Fully connected layer CNN ends with one or more fully connected layers that produce non-spatial output. The fully connected layers use the features arriving from all the previous layers for classifying the input image into vary classes. The number of

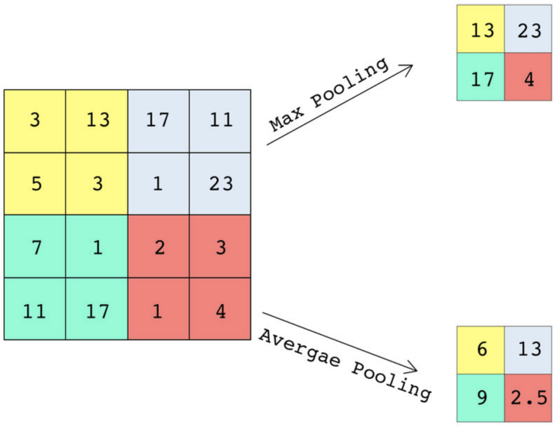


Figure 3.2: *Example of max pooling and average pooling operations. In this example a 4x4 image is downsampled to a 2x2 by taking the maximum value or the average value of each sub-region.*

neurons of the last fully connected layer equals the number of the classes. Figure 3.3 shows the overall CNN structure.

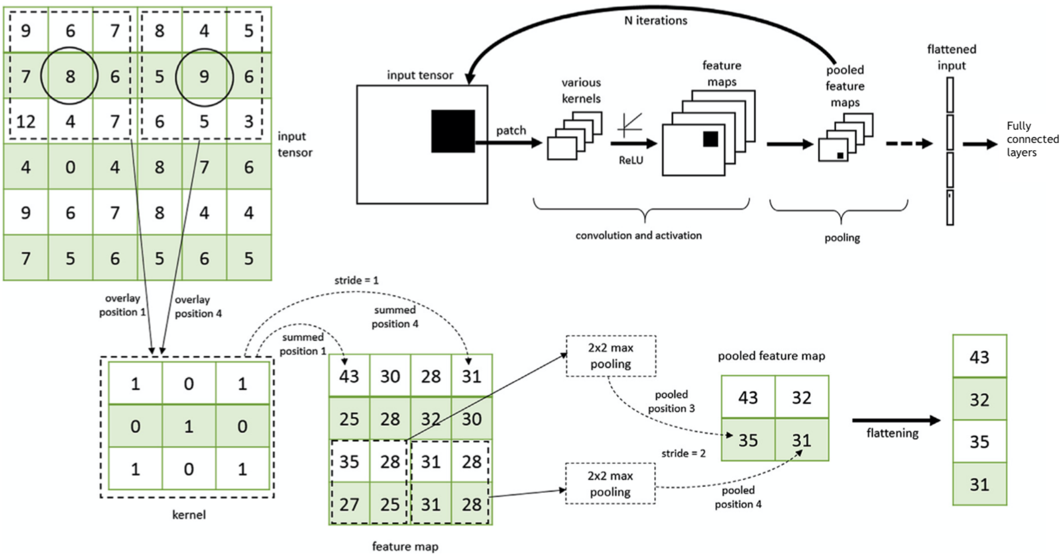


Figure 3.3: *The figure summarizes the structure and functioning of a general CNN. The CNN has a number of convolution and pooling layers before flattening and input to the fully connected layers. The figure shows a schematic representation of a convolution using a 3 x 3 kernel with a stride of 1. The weighted sum of the kernel for the 3 x 3 input tensor creates a single representative value in the feature map. Multiple feature maps are produced by different kernels. Max pooling is used with a 2 x 2 array, and a stride of 2. The final feature map is flattened to be processed as input for the dense layer.*

3.2 Proposed approach

The proposed method to I frame selection exploits the power of deep learning and relies on the transfer learning strategy. Three different CNNs pre-trained on natural images are used as both feature extractor and classifier to correctly classify frames as:

- **I**, frames that clearly show the hyaline cartilage, delimited by the two regular, sharp, and continuous hyperechoic interfaces (Figure **??**).
- **NI**, frames that do not clearly show the chondrosynovial and the osteochondral interfaces (Figure 1.4b).

For a fair comparison all the CNN models were pre-trained on the ImageNet dataset. The selected CNNs make their decision based on highly abstracted features difficult for humans to understand. In this work it is proposed a visual explanation called grad-CAM, which enables the systems to highlight the important regions in the frame where CNNs focus while making the final image classification. The workflow of the proposed approach is shown in Figure 3.4

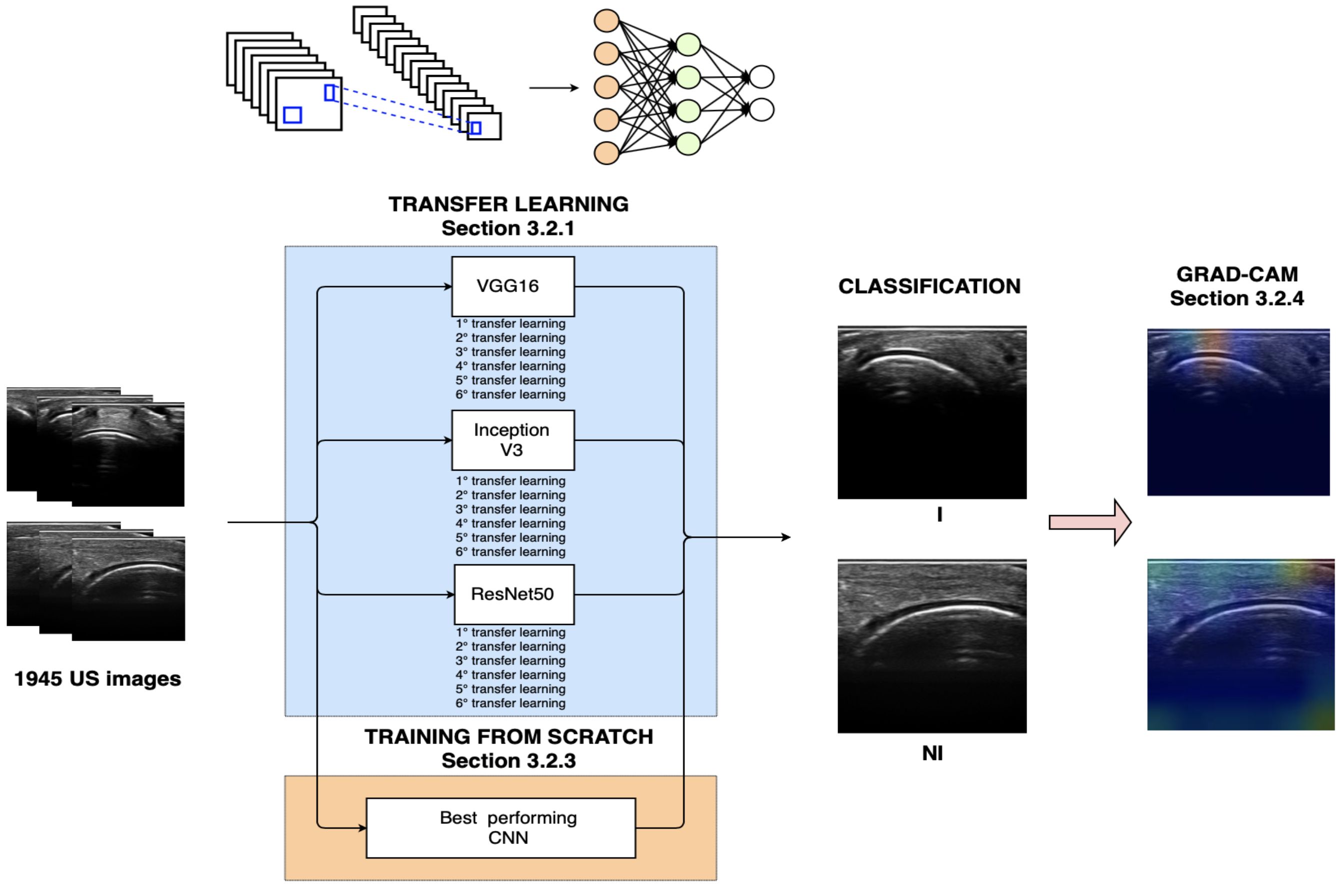


Figure 3.4: *Workflow of the proposed approach to I frame selection in US metacarpal head scanning. The frames are classified as I: informative and NI: not informative. The approach leans on the transfer learning strategy. Six different transfer learning are adopted for each selected CNN. For the best performing CNN the the training from scratch is actuated. Finally grad-CAM is applied for interpretability purposes.*

3.2.1 Transfer learning strategy

Data dependence is one of the most serious problem in DL. In particular, CNNs have a very strong dependence on massive training data, because they need a large amount of data to understand the latent patterns of images. Moreover training a CNN from scratch requires an extensive computational and memory resources. A promising alternative to training CNNs from scratch is transfer learning. Inspired by humans capability in applying learned knowledge to solve new problems faster and with better solutions, transfer learning aims to exploit and to transfer the knowledge or weights of a neural network trained on a large dataset for new tasks. Practically, it generalizes the knowledge (features, weights) of an existing solution to a new problem, leading to promising results also when the new task has significantly less data.

As reported in [53], the formal definition for deep transfer learning involves the concepts of domain and task. A domain can be represented by $\mathcal{D} = \{\mathcal{X}, P(\mathrm{X})\}$, which contains two parts: the feature space $\mathcal{X}$ and the edge probability distribution $P(\mathrm{X})$ where $\mathrm{X} = \{x_1,...,x_n\} \in \mathcal{X}$. A task can be represented by $\mathcal{T}$=$\{y$, *f(x)*$\}$. It consists of two parts: label space y and target prediction function *f(x)*. *f(x)* can also be regarded as a conditional probability function $P(y|x)$. Then, the transfer learning can be formal defined as follows: given a source domain $\mathcal{D}_s$ and learning task $\mathcal{T}_s$, a target domain $\mathcal{D}_t$ and learning task $\mathcal{T}_t$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_t$ using the knowledge in $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{D}_s \neq \mathcal{D}_t$, and/or $\mathcal{T}_s \neq \mathcal{T}_t$. In addition, in the most case, the size of $\mathcal{D}_s$ is much larger than the size of $\mathcal{D}_t$. The learning process of the transfer learning is illustrated in Figure 3.5.

In this work three CNNs pre-trained on ImageNet were investigated: VGG16, Inception V3 and ResNet50. These architecture have been chosen because they let to investigate three different levels of depth for the learned feature extraction; furthermore they allow to learn complex and fine-levels features while making the train convergence easier and faster. Moreover, as demonstrated in [54], they are perfectly suitable for network-based transfer learning and they allow to investigate different level of depth.

Network-based transfer learning refers to the reuse the partial network pre-trained in the source domain, including its network structure and connection parameters, transfer it to be a part of the deep neural network which is used in the target domain [53].

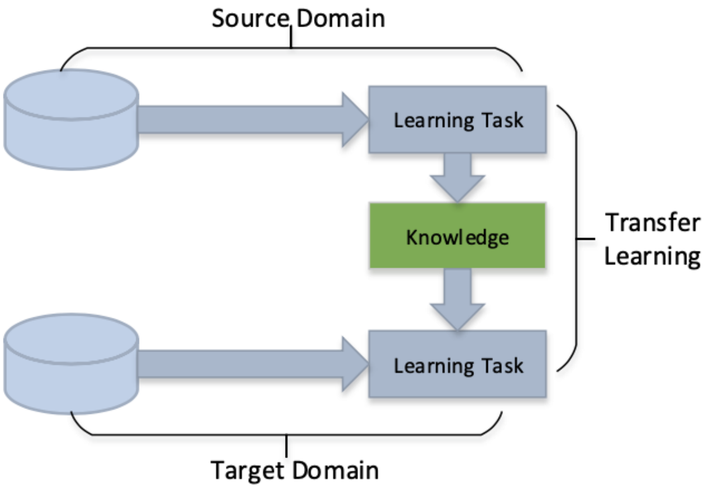


Figure 3.5: *Graphical representation of the transfer learning approach of previous works. The knowledge (features, weights) that a model has learned from a task (i.e., natural image classification) where a lot of labeled training data are available (source domain) is exploited and transferred to another task, such as medical image classification, with less data (target domain)*

In other words, it consists of copying the first n layers of a pre-trained network to the first n layer of the target network. The copied n layers are said frozen since they are not updated during the training process of the new task.

The remaining layers of the target network can be randomly initialized or fine-tuned and trained toward the target task. Fine-tuning the remaining layers involves initializing the weights with the transferred features and to update them during the training.

In most of the network-based transfer learning approaches for the automatic frame selection [45, 46] the fully connected layers are fine-tuned, while the convolutional layers are frozen and used as feature extractors. Freezing the first layers is done due to the fact that they refer to general features. Figure 3.6 displays the just described approach.

One of the main aspect of CNNs is that features extracted by the first layers are not specific to particular dataset, but general and applicable to many dataset. Features computed by the last layer of a trained network must depend greatly on the chosen dataset and task. Thus, these last layer features are specific. If first layer features are general and last layer features are specific, then there must be a transition from general

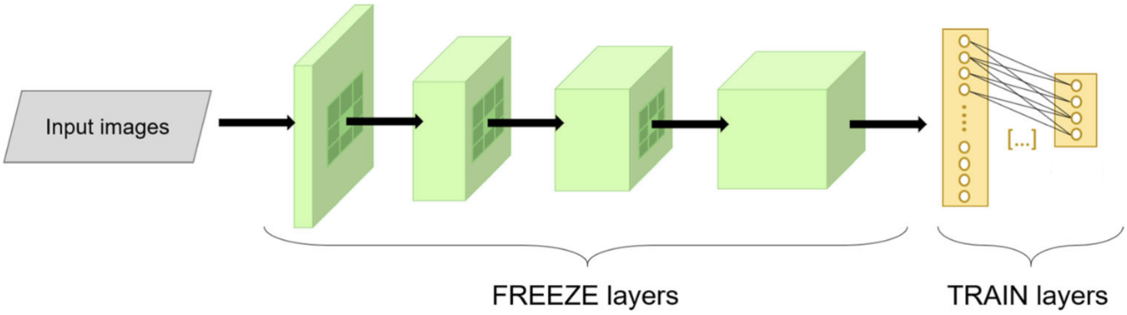


Figure 3.6: *Graphical representation of the transfer learning technique. The convolutional layers are frozen and used as features extractor, while the fully connected layers are fine-tuned. This figure is a courtesy of [46]*

to specific somewhere in the network.

Identifying the transition from general to specific can lead to time saving and higher performances. Since general features can be applied to many dataset, avoiding their fine-tuning can save time; on the other hand, fine-tuning specific features layers can lead to higher performances.

In this work, in order to identify the general-specific feature transition, 6 different network-based transfer learning are applied for each CNN. In each network-based transfer learning approach, the remaining convolutional layers are fine-tuned.

In Table 3.1 are listed the investigated CNNs.

Table 3.1: *Tested CNNs and the top-1 and top-5 accuracies achieved on the ImageNet dataset. These accuracies refer to the fractions of test images for which the correct label is the first (top-1) or among the five labels (top-5) considered most probable by the model, respectively.*

Model	**Top-1 accuracy**	**Top-5 accuracy**
VGG16	71.3%	90.1%
Inception V3	77.9%	93.7%
ResNet50	74.9%	92.1%

The tested CNN architectures are hereafter briefly described to highlight their main peculiarities.

VGG16 VGG16 was proposed by the Oxford's Visual Geometry Group (VGG) in the context of Large Scale Visual Recognition Challenge (ILSVRC) in 2014. VGG16 improved the performance of previously proposed deep networks (e.g., AlexNet) by replacing large-sized kernel filters with stacked kernels with dimension 3x3 pixels.

VGG16 has a uniform (serial) architecture with 13 convolutional and 5 (downsampling) max pooling layers, followed by 3 fully-connected layers. The convolution stride is fixed to 1 pixel, while max-pooling is performed over a 22 pixel window, with a stride of 2. The number of channels starts from 64 in the first layer and then increases by a factor of 2 after each max-pooling layer, until it reaches 512.

In the adopted transfer learning strategy, the three fully connected layers of the VGG16 were modified: 1024 neurons in the first layer with a $ReLU$ activation function, 512 in the second layer activated by a *tanh* and 2 in the last layer since it is a binary classification problem. The VGG16 ends with a softmax activation function which returns a probability distribution over the target classes.

Six network-based fine-tuning approaches were accomplished at different heights of the architecture and they included the complete fine-tuning of the VGG16 structure (VGG16 0), four intermediate fine-tuning (VGG16 1, VGG16 2, VGG16 3, VGG16 4), and the fine-tuning of just the classifier of the VGG16 (VGG16 5).

The VGG16 architecture and transfer learning approaches are shown in Figure 3.7

Inception V3 The winner of ILSVRC 2014 competition was GoogLeNet (i.e., Inception V1) developed by Google LLC.

The innovative idea of GoogLeNet was the introduction of the Inception module, the building block of the Inception V3 architecture. In comparison to GoogLeNet, Inception V3 mainly focuses on burning less computational power resulting in more computationally efficient network.

The Inception V3 contains 11 Inception modules for a total of 48 layers. An Inception module executes simultaneously on the same input different convolutional operations, whose kernels have dimensions 1x1, 3x3, and 5x5. The different activation maps coming form the same input are then concatenated. The Inception V3 processes the image at varying scale when it passes through the CNN modules.

For the purpose of binary classification (I and NI), the last fully connected layer was modified by reducing the number of neurons from 1000 to 2. The Incpetion V3 ends with a softmax activation layer.

To identify the generic to specific feature transition six different transfer learning proposals were applied: the complete fine-tuning of the Inception V3 (Inception V3 0), the the fine-tuning the Inception V3 classifier (Inception V3 5), and four intermediary fine-tuning (Inception V3 1, Inception V3 2, Inception V3 3 and Inception V3 5).

The architecture of Incpetion V3 and the details of the transfer learning approaches are displayed in Figure 3.8

ResNet50 ResNet architectures won the first place on the ILSVRC 2015 classification task [55]. They address the problem of vanishing gradients during the training of deep neural networks.

The breakthorugh of ResNets is the introduction of the residual units, that allow the training of ultra deep neural networks without encountering the issue of saturation performances.

The residuals units are the building blocks of the ResNet50. The skip connections are the advance introduced by the residual units. They skip some layers in the neural network and feeds the output of one layer as the input to the next layers. The architecture of ResNet50 consists of 5 stages and the number of building block varies in each stage (3, 4, 6, and 3 respectively). Each building block is made of three convolutional kernels with skip connections. In the fully connected layer the number of neurons is reduced to two for the I and NI classification task. The ResNet50 ends with a softmax layer too.

Six different transfer learning were carried out to identify the specific features extractor layers. As in the previous two architectures the six adopeted strategies included the complete fine-tuning of the network (ResNet50 0), four half-way fine-tuning (ResNet50 1, ResNet50 2, ResNet50 3, ResNet50 4) and the ResNet50 classifier fine-tuning (ResNet50 5).

A detailed representation of the network and transfer learning strategy is in Figure 3.9

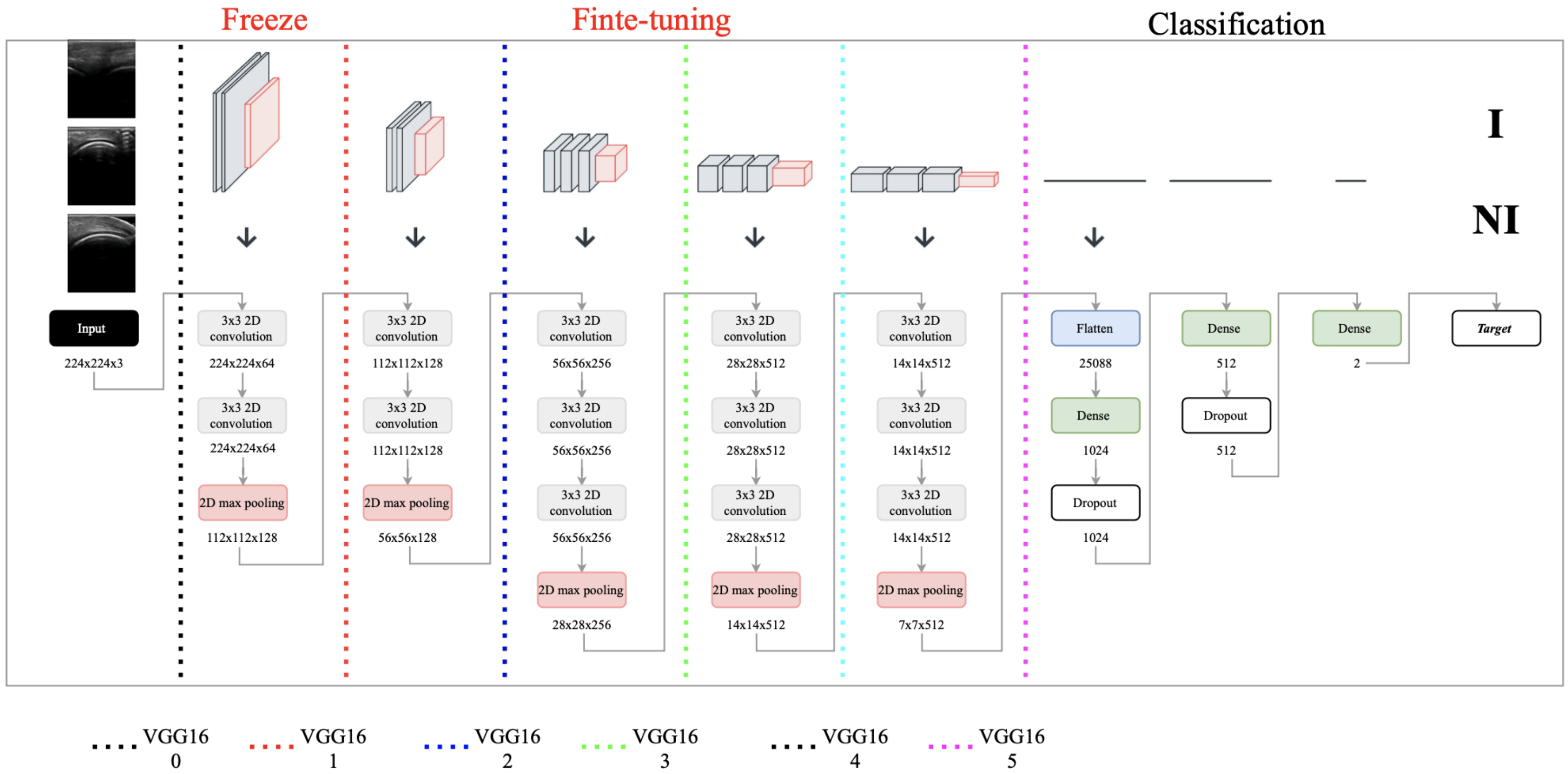


Figure 3.7: *The figure shows the details of the VGG16 structure used in this work. The six dot lines represent the six different adopted transfer learning. Considering the transfer learning used for the VGG16 1 (red dot line): the first convolutional layers are frozen, while the reaming ones are fine-tuned.*

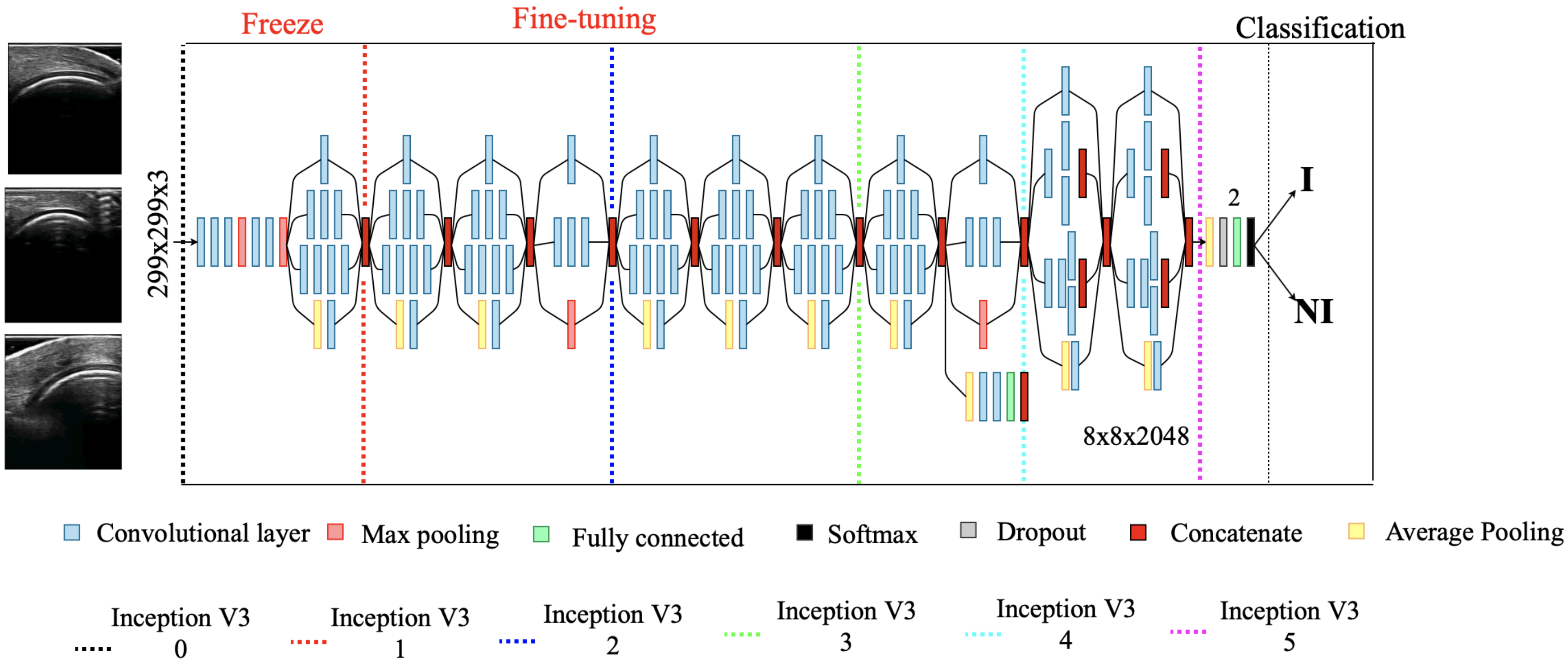


Figure 3.8: *The figure exposes a detailed structure of the Inception V3 used in this work. The six dot lines represent the six different adopted transfer learning. Considering the transfer learning used for the Inception V3 1 (red dot line): the left part is the frozen architecture, while the reaming layers are fine-tuned.*

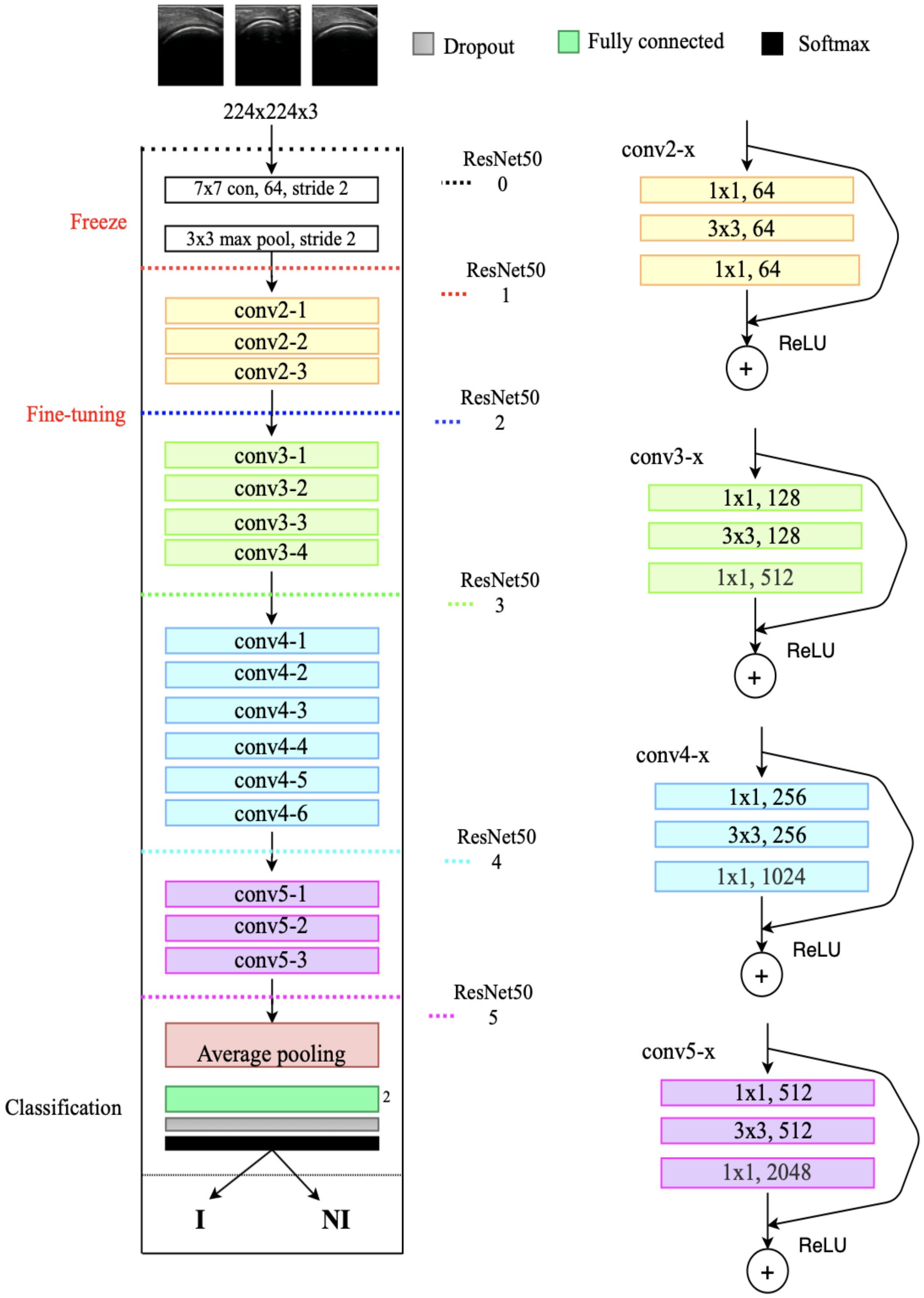


Figure 3.9: *The figure shows the structure of the ResNet50 used in this work. The six dot lines represent the six different adopted transfer learning. Considering the transfer learning used for the ResNet50 1 (red dot line): the left part is the frozen one, while the reaming layers are fine-tuned.*

3.2.2 Training settings

Deep learning involves optimization, which refers to the task of minimizing an objective or loss function *J(w)* by alterating w (model's parameters). The exploited optimizer is a variant of the stochastic gradient descend (SGD): the mini-batch SGD. This optimizer splits the training set into small batches. In this way model's parameters update frequently avoiding local minimum for a more robust convergence. Using the mini-batch SGD optimizer, every model's parameter (w) after each mini-batch of n training examples is computed in this way:

$$w = w - \alpha \nabla_w J(x^{(i:i+n)}, y^{(i:i+n)}) \tag{3.2}$$

where J(θ) is the objective function, α the learning rate, x the training examples and y the corresponding labels.

During the training the binary cross entropy was minimized and it is defined as:

$$CE = -\sum_{i}^{C} t_i log(p_i) \tag{3.3}$$

being i the sample $\in C = \{I, NI\}$, t_i the target, and p_i the given prediction.

3.2.3 Frame classification using training from scratch

To verify the efficiency of the transfer learning strategy, an additional experiment was carried out. For the best performing architecture it is implemented a trained from scratch. Learning from scratch means that neural network does not have any stored knowledge for solving the new task and features are directly learned during the training procedure.

Weights were initialized with Glorot initialization [56] that involves the initialization of each weight with a small Gaussian value with 0 mean and variance based on incoming (n_{in}) and outcoming (n_{out}) weights, as described in the formula:

$$Var(w) = \frac{2}{n_{in} + n_{out}} \tag{3.4}$$

where *Var(w)* indicates the variance of the initialized weight.

3.2.4 Grad-CAM

DL models provide the possibility to find solutions for many problems related to image classification, object detection, and image segmentation. Despite the higher-grade performances that they enable, they remain mostly black boxes. In particular, in CNNs, as the input image propagates toward deeper layers, the level of abstractness of the extracted features increases, making them hard to interpret. This aspect is crucial because if on one hand CNNs offer the possibility to reach superior performances, on the other hand they don't provide interpretability. In other words, when a CNN fails in solving a specific task, there are not direct explanations for the incoherent output, leaving the user to attempt to find out a reasonable justification on why the system did what it did.

Typically exists a trade off between interpretability and performances. The classical rule-based is: simple CNNs are underachievement. By using deeper and more complex CNNs interpretability is sacrificed for greater performances.

In the clinical practice understating and justifying a specific result is of primary importance. Thus, interpretability represents the key element for the introduction of deep models in the clinical daily life.

Grad-CAM is a visual explanation method able to add interpretability and faithfulness to CNNs. It enables to generate heatmaps for a specific class, highlighting the class-like different parts of the image. The grad-CAM algorithm is very simple. An overview workflow is represented in Figure 3.10.

Considering that convolutional layers retain spatial information which is lost in the fully connected layers, is straightforward that the last convolutional layer contains the best high-levels semantics and detailed spatial information. The neurons in the fully connected layers look for semantic class-specific information in the image. Grad-CAM exploits the gradient of the score for class c, y^c, with respect to the feature map activation A^k of the last convolutional layer. These gradients flowing back are global average pooled (pool size equals the input size, so that the average of the entire input is computed as the output value) over the width and the height dimensions (indexed

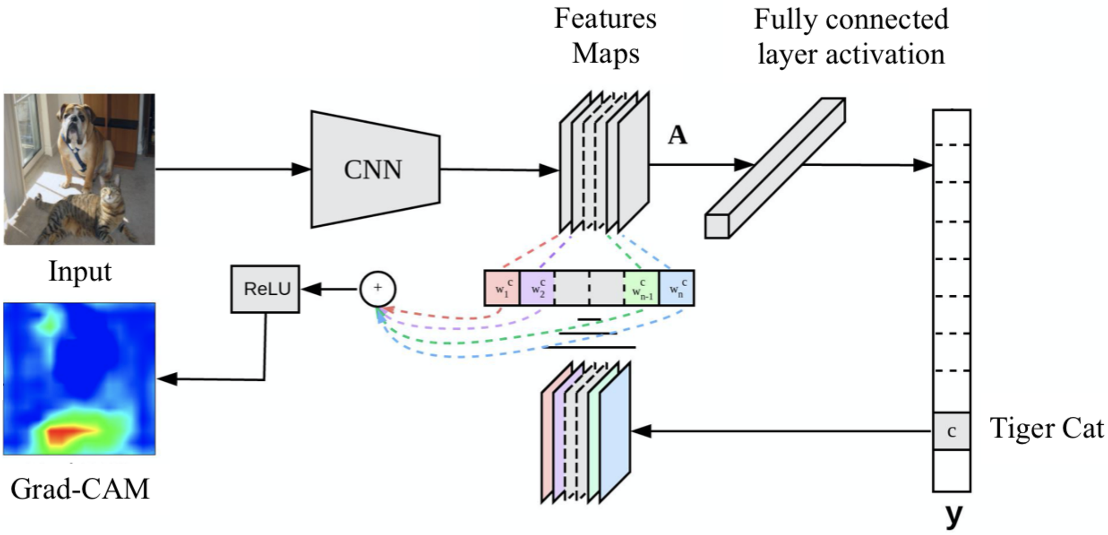


Figure 3.10: *Overview of grad-CAM. Given an image and a class of interest, in this case tiger cat, the algorithm first forward propagate the image through the CNN to obtain the score of the of the category. Then gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model looks when making the particular decision.*

by i and j) to obtain the neuron importance weights w_k^c.

$$w_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \qquad \underbrace{\frac{\delta y^c}{\delta A_{ij}^k}}_{\text{gradients via backprop}} \tag{3.5}$$

In order to obtain the the class-discriminative localization map Grad-CAM $L^c_{grad-CAM}$ (represented in blue in Figure 3.10) is performed a weighted combination of forward activation, followed by a *ReLU*.

$$L^c_{grad-CAM} = ReLU\left(\sum_k w_k^c A^k\right) \tag{3.6}$$

The choice of using the *ReLU* activation function is guided by the need of highlighting only the features that have a positive influence on the class of interest c to increase the corresponding score y^c [51]. Finally grad-CAM results in a spatial map of how intensely the input image activates a certain class.

In this thesis the employment of grad-CAM aims to built a more transparent automatic frame selection system that exploits the advantages of CNNs.

Chapter 4

EXPERIMENTAL PROTOCOL

Synopsis In this Chapter, the experimental protocol of the proposed method is described. In particular, Section 4.1 reports the proceeding for the arrangement of the training, validation and testing set. In Section 4.2, all the choices relative to the tuning parameters for the training procedure of the proposed strategy are described . Section 4.3 describes the performance metrics for the evaluation of the best CNN and for the CNN trained from scratch. Finally Section 4.4 indicates the technical specifications of the hardware and software used for the implementation of the algorithm.

4.1 Dataset

In this study 1945 US frames, belonging to 40 healthy subjects, were analyzed. The images were acquired using MyLabClassC (Esaote, Genoa, Italy) equipped with a very high frequency bradband linear probe (10-22 MHz). A grey-scale standard setting was adopted with a B mode and frequency of 22 MHz. Metacarpal head from the 2^{nd} to the 5^{th} digit of both hands was scanned.

For the purpose of building the dataset, the clinician consciously performed wrong acquisitions committing the most common mistakes in US scanning.

The acquired frames were manually labeled by the clinician. In each subject there is a balance number of I and NI frames.

A leave 4-subjects out cross validation was performed to test the fine-tuned networks: it consists of 10 iterations in which in each iter 4 subjects are left out as test patients, while the remaining subjects are randomly selected for the training (28 subjects) and validation (8 subjects) set. Finally, 10 different models are shaped. At each iteration the 4 subjects for the testing set are recast.For comparison purposes the ten models were the same for all the six transfer learning configurations of each CNN. Figure 4.1 show in details the leave 4-subjects out cross validation.

Thanks to this division set, images belonging to the same patient constitute only the training, validation or testing set. Maintaining the subject division is crucial; because in the testing set there must be images never seen by the neural network. Since metacarpal head sonograms comprehend characteristics shared by all the subjects, and other specific for single individuals; maintaining the subject division will create a testing set with images having specific features never seen by the network.

For fair comparison, the leave 4-subjects out cross validation iterated 10 times was used for testing purposes also when testing the performances of the CNN trained from scratch. In Figure 4.2 are displayed samples of images belonging to the dataset.

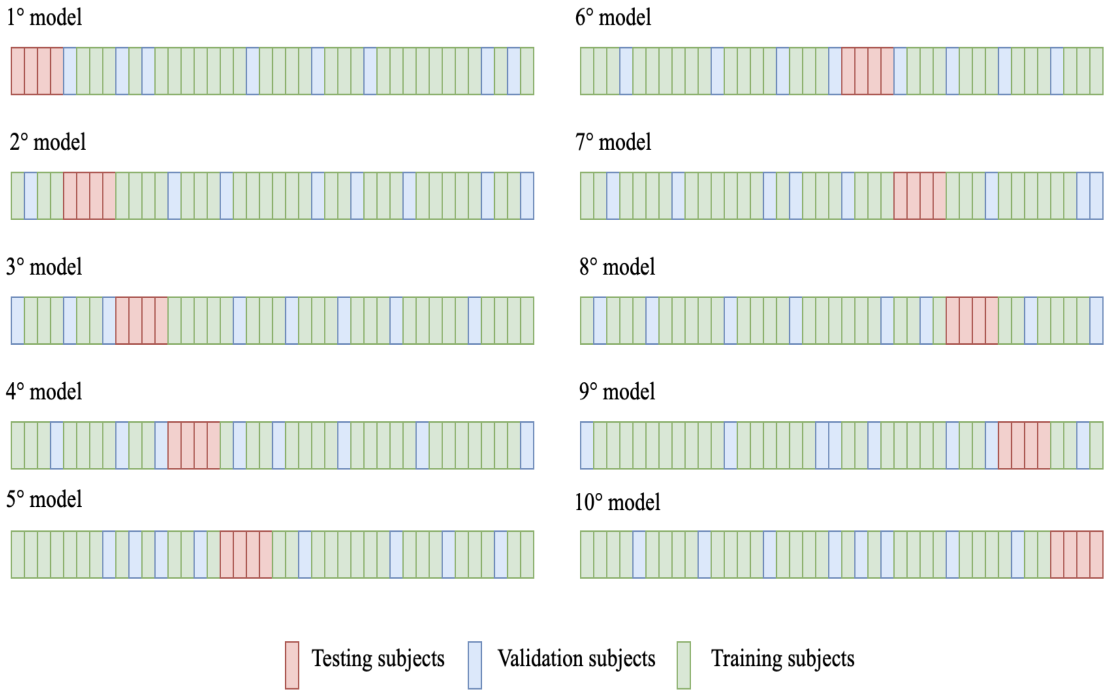
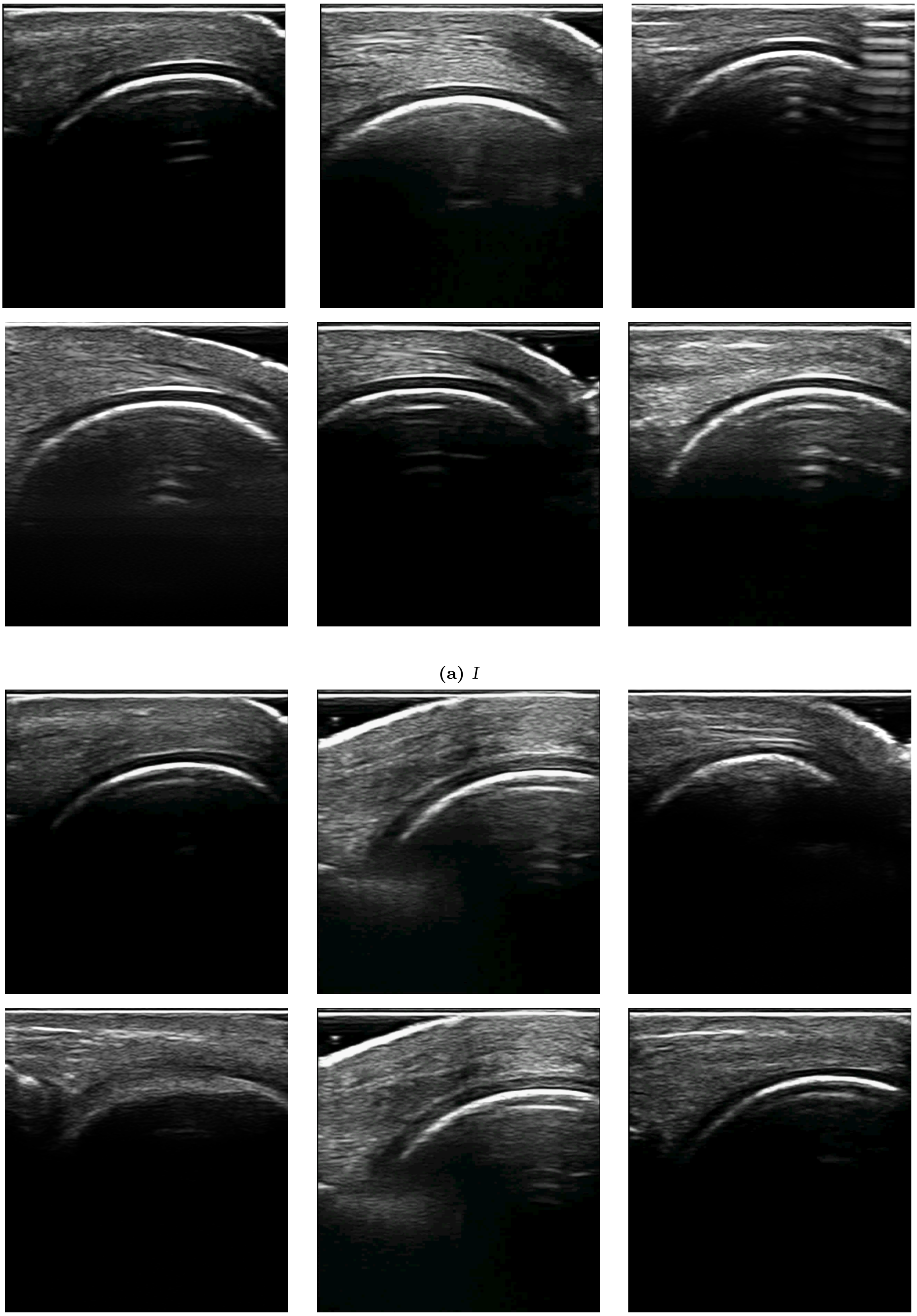


Figure 4.1: *The figure shows the leave 4-subjects out cross validation iterated ten times. The eight subjects for the validation set were randomly selected for each of the 10 models. For repeatability and comparison purpose in each transfer learning configuration the 10 models were the same.*



(a) *I*

(b) *NI*

Figure 4.2: *Samples of metacarpal head frames belonging to the dataset. In (a) are displayed informative frames, while in (b) the not informative ones.*

4.2 Fine-tuning parameters

To extract the learned features from each metacarpal head sonograms with the architecture described in Section 3.2.1, all the frames were resized to match the input size of the investigated CNN architecture. In details, to match the input size of the VGG16 and ResNet50 the metacarpal frames were resized to 224x224x3, while for the Inception V3 the images were resized to 299x299x3. Since sonograms are gray scale images, they have only one channel. In order to fit the input size of the networks, copies of the same gray scale images in the 3 different channels were made. Before the training the images were preprocessed by removing the intensity mean.

For the mini-batch SGD the batch size is set to 64 as a balance between training speed and gradient convergence. The batch size is a hyperparameter that defines the number of samples to work through before updating the internal model parameters.

The number of epochs, that defines the number of times that the learning algorithm will work through the entire training dataset, is set to 100.

The performance of the mini-batch SGD, described in Section 3.2.2, depends critically on how the learning rate (α) is tuned. α determines the step size of each iteration while moving toward a minimum of the loss function (cross entropy).

In setting the learning rate, there is a trade-off between the rate of convergence and overshooting. While the descent direction is usually determined from the gradient of the loss function, the learning rate determines how big a step is taken in that direction. A too high learning rate will make the learning jump over the minima but a too low learning rate will either take too long to converge or get stuck in an undesirable local minimum.

As described in Section 3.2.1, when the network is fine-tuned, the weights are initialized with the transferred features of ImageNet. Due to this initialization and to the assumption that the gradient of the cross entropy is already in a good position (coming from ImageNet), the learning rate is set to 10^{-4} in order to not move too far from the favorable area of where the gradient descent starts.

The momentum is another tuning parameter that speeds up the descent of the gradient along the correct direction, avoiding local minimum. In this thesis the momentum is set to 0.9.

For a fair comparison the same values were used for the CNN trained form scratch.

4.3 Data analysis

The classification performances of each tested CNN model were evaluated with respect to the manual annotation performed by the clinician, considered as the ground truth. A set of metrics commonly used to evaluate the performance of a binary classification is employed [57]. Defining first a series of terms:

- *True Positive (TP)*: number of predicted positives correctly classified: number of I frames correct classified as I frames.

- *True Negative (TN)*: number of predicted negative correctly classified as number of NI frames correctly classified as NI.

- *False Positive (FP)*: number of predicted positives incorrectly classified like number of I frames misclassified as NI.

- *False Negative (FN)*: number of predicted negative incorrectly classified as the number of NI frames misidentified as I ones.

The metrics used for the evaluation were:

- Class-specific precision ($Precision_{i,i \in [I,NI]}$), defined as

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{4.1}$$

- Class-specific recall ($Recall_{i,i \in [I,NI]}$), defined as

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{4.2}$$

- Class-specific F1-score ($F1\text{-}score_{i,i \in [I,NI]}$), defined as

$$F1 - score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4.3}$$

The area under the curve (AUC) of the mean Receiver Operating Characteristic (ROC) was used to evaluate the classification performances of the six transfer learning configurations of the three CNNs because unlike accuracy, it describes the discriminatory power of a classifier independently of the class distribution [58].

For the best CNN with the highest AUC, it was performed the training from scratch.

4.4 Technical specification

The overall algorithm was implemented with Keras [1]. Indeed, all the CNN models and weights were downloaded from the Keras library.

Experiments were performed using Google Colaboratory [2]: a free GPU cloud platform based on Jupyter notebook environment that supports 14858 MB of free GPU, 12.4 GB of free RAM and a processor size of 900.6 MB.

[1] https://keras.io

[2] https://colab.research.google.com/notebooks/intro.ipynb#recent=true

Chapter 5

RESULTS

Synopsis In this Chapter are reported the results of the applied method described in Chapter 3, in terms of performances metrics described in Chapter 4. Section 5.1 documents the results of the best fine-tuning configurations of each selected CNN. In Section 5.2 the results of the CNN trained from scratch are reported. The grad-CAM results are described in Section 5.3.

5.1 Fine-tuning results

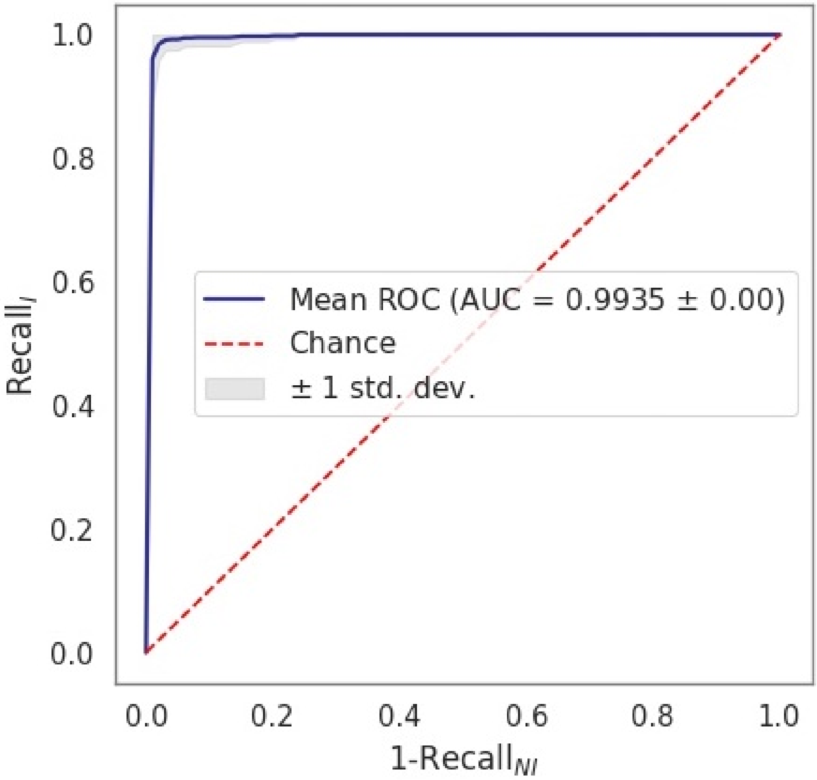
Among the six transfer learned configurations of the selected CNNs (VGG16, Inception V3 and ResNet50) only the results of the best performing networks are reported.

Figure 5.1 displays the mean ROC and the AUC results of the VGG16 0, Inception V3 0 and ResNet50 0, resulted to be the fine-tuned configurations with the highest discriminative power (AUC) for each CNN.

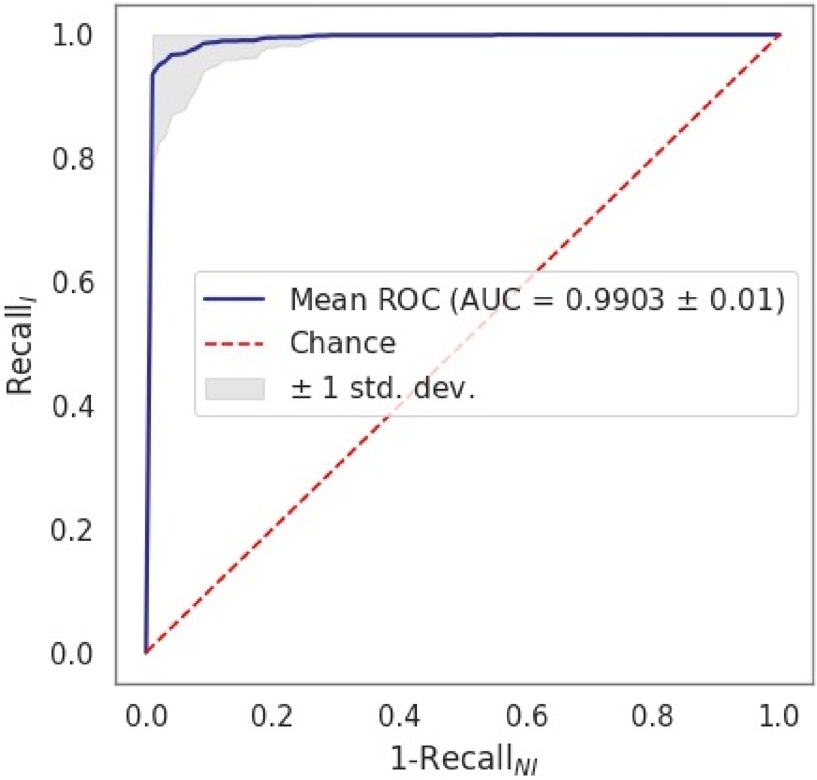
In particular, VGG16 0, Inception V3 0 and ResNet50 0 are among the six adopted transfer learning approaches, the ones in which the fine-tuning strategy involves the overall architecture. A more detailed visualization of the three fine-tuning architectures is in Figure 3.7 for the VGG16 0, in Figure 3.8 for the Incpetion V3 0, and in Figure 3.9 for the ResNet50 0.

The VGG16 0 in Figure 5.1a, has the highest discriminatory power with an AUC of 0.9935, followed by the ResNet50 0 (Figure 5.1c) with an AUC of 0.9921 and at the

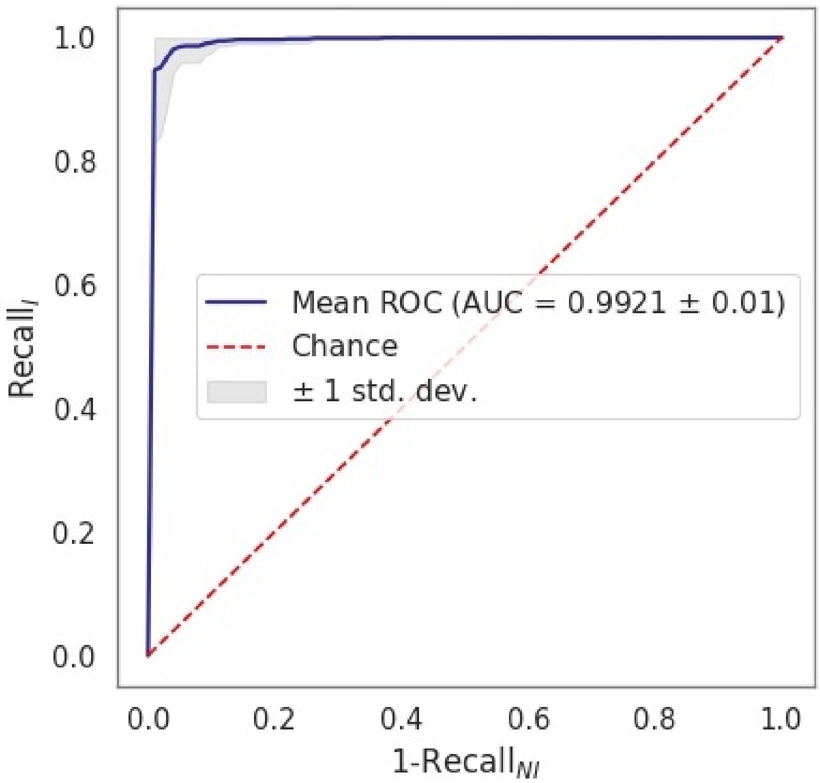
Incpetion V3 0 (Figure 5.1b) with an AUC of 0.9903.



(a) *ROC VGG16 0*



(b) *ROC Inception V3 0*



(c) *ROC ResNet50 0*

Figure 5.1: *The figure shows the mean ROC and the corresponding AUC results for the best performing fine-tuned configuration of the three selected CNNs. In (a) the ROC and AUC value for the VGG16 0, in (b) for the Inception V3 0 and in (c) for the ResNet50 0.*

The performances metrics of the VGG16 0, the Inception V3 0, and the ResNet50 0, described in Section 4.3, are reported in Table 5.1. The values reported in the table are the average performances among the ten models.

The VGG16 0 achieves a value of 0.99 in terms of precision, recall and F1-score for the classification of I frames. The same value is preserved for the recall and F1-score of the NI frames, despite the precision which has a value of 0.98. The VGG16 completely fine-tuned results to be the superior respect to the ResNet50 0 and the Incpetion V3 0.

Comparing the ResNet50 0 and the Inception V3 0; the first one shows higher average metrics performances respect to the Inception V3 0. Nevertheless, Inception V3 0 exceeds in terms of precision for the classification of I frames the ResNet50 0 with a value of 0.99. To resume the VGG16 0 is the best acting CNN followed by the ResNet50 0 and the Inception V3 0.

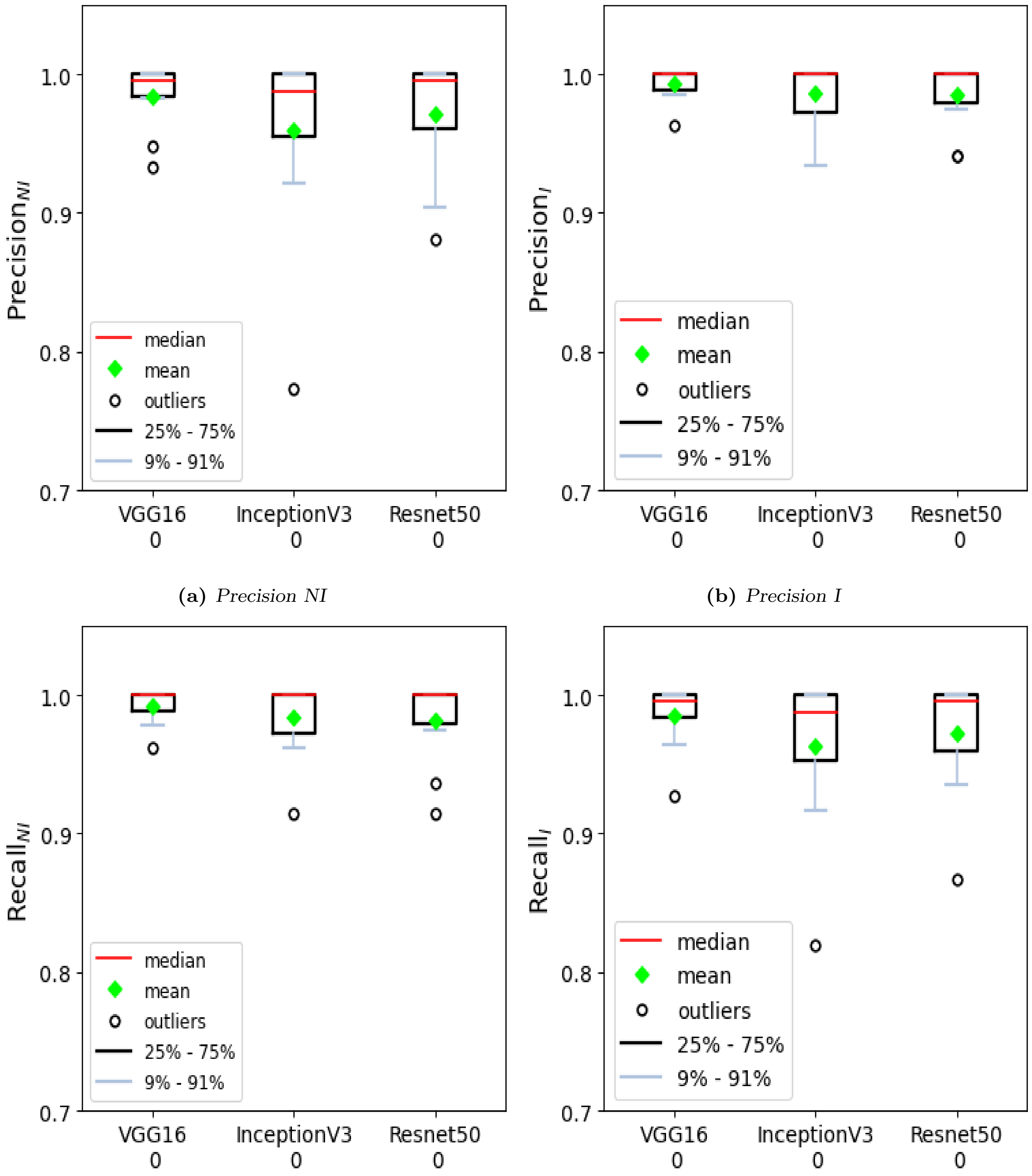
Table 5.1: *Metrics performances of the VGG16 0, Inception V3 0 and ResNet50 0. The value inside the round brackets represents the standard deviation. Performances are reported in terms of class-specific precision, recall, and F1-score.* ***NI****: non-informative frame,* ***I****: informative frame,* ***avg****: average.*

	Precision$_i$	**Recall$_i$**	**F1-score$_i$**
VGG16 0			
NI	**0.98 (0.02)**	**0.99 (0.01)**	**0.99 (0.01)**
I	**0.99 (0.01)**	**0.99 (0.02)**	**0.99 (0.01)**
avg	**0.99 (0.01)**	**0.99 (0.01)**	**0.99 (0.01)**
Inception V3 0			
NI	0.96 (0.07)	0.98 (0.03)	0.97 (0.05)
I	0.99 (0.02)	0.96 (0.05)	0.97 (0.05)
avg	0.97 (0.04)	0.97 (0.04)	0.97 (0.04)
ResNet50 0			
NI	0.97 (0.04)	0.98 (0.03)	0.98 (0.03)
I	0.98 (0.02)	0.97 (0.04)	0.98 (0.03)
avg	0.98(0.03)	0.98(0.03)	0.98(0.03)

Boxplots were used to compare the robustness of each CNN among the ten leave 4-subjects out cross validation models .

Figure 5.2a and Figure 5.2b compare the behavior of the three best fine-tuning architectures in terms of class specific precision.

The outliers of the VGG16 0 are above the 0.9 while for the ResNet50 0 and Inception V3 0 they fall below the same value. Moreover the boxes for the VGG16 0 are narrower compared to the boxes of the other two CNNs. A similar behaviour can be observed also in Figure 5.2c and Figure 5.2d for the recall performance and in Figure 5.2e and Figure 5.2f for F1-score metric.

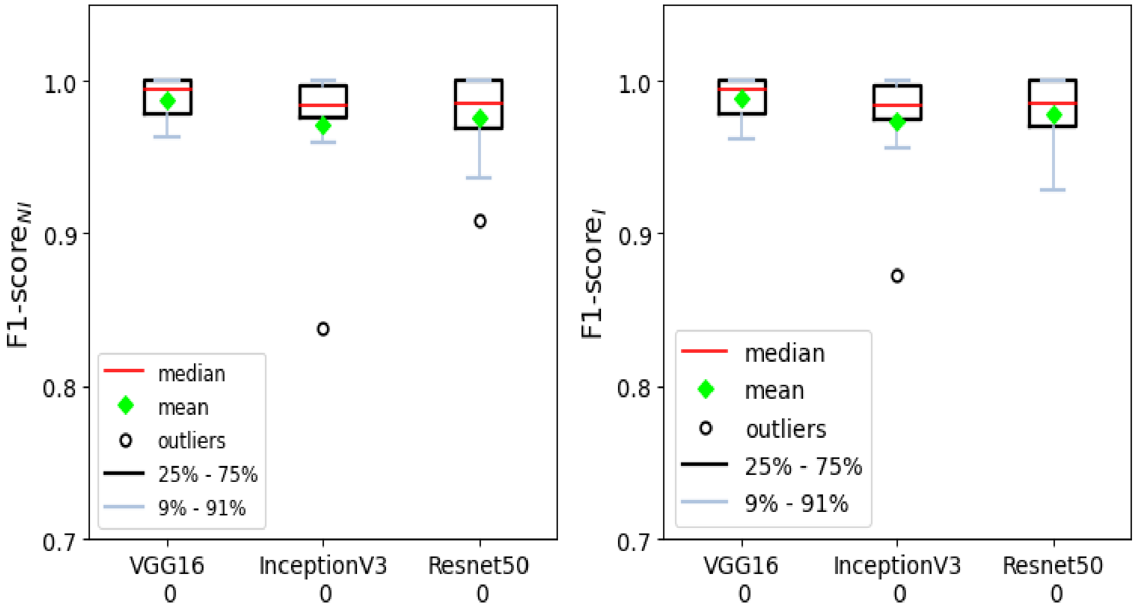


(a) *Precision NI*

(b) *Precision I*

(c) *Recall NI*

(d) *Recall I*



(e) *F1-score NI* **(f)** *F1-score I*

Figure 5.2: *Boxplots of classification precision, recall and F1-score for the NI (a,b,c) and I (d,e,f) frames.*

5.2 Results of the CNN trained from scratch

Since the VGG16 completely fine-tuned is architecture with the highest AUC, the training from scratch of this architecture was performed.

The mean ROC of the VGG16 trained from scratch is displayed in Figure 5.3. VGG16 trained from scratch reaches an AUC of 0.9904, lower than the VGG16 completely fine-tuned, confirming the efficiency of the transfer learned approach.

The classification performances of the VGG16 trained from scratch are reported in Table 5.2 in terms of class-specific precision, recall and F1-score. The precision, recall and an F1-score for the classification of I frames of the VGG16 trained from scratch are respectively 0.98, 0.96 and 0.97; lower than the results obtained with the VGG16 0 pre-trained on ImageNet.

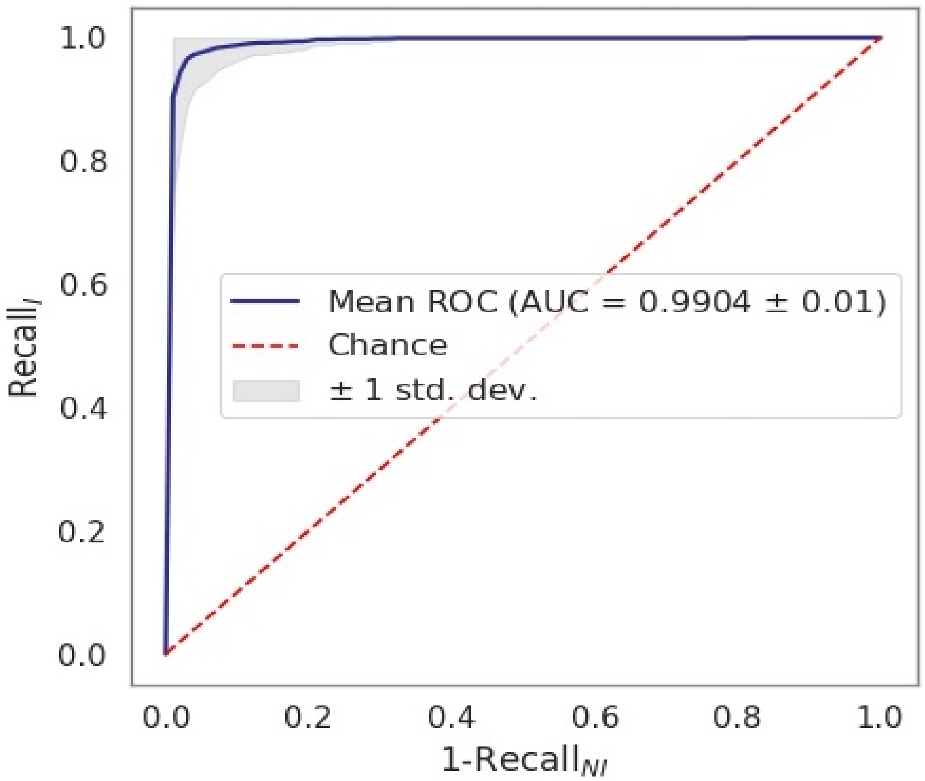


Figure 5.3: *The figure shows the mean ROC and the corresponding AUC of the VGG16 trained from scratch.*

Table 5.2: *Metrics performances of the VGG16 trained from scratch. The values inside the round brackets represent the standard deviation. Performances are reported in terms of class-specific precision, recall, and F1-score.* ***NI****: non-informative frame,* ***I****: informative frame,* **avg***: average.*

	Precision$_i$	**Recall**$_i$	**F1-score**$_i$
VGG16 from scratch			
NI	0.96(0.07)	0.98(0.02)	0.97(0.04)
I	0.98(0.02)	0.96(0.06)	0.97(0.04)
avg	0.97(0.04)	0.97(0.04)	0.97(0.04)

5.3 Grad-CAM results

The grad-CAM results are displayed in Table 5.3 and in Table 5.4 for the classification of respectively I and NI frames.

The red areas inside the images refer to the areas where the CNN focuses before making the final decision. It highlights the class-discriminative regions.

From Table 5.3 is possible to understand that VGG16 0 perfectly classifies the I frames. Moreover the red areas are always centered in the chondrosynovial interface. From the same table the ResNet50 0 misclassifies the 4° and 5° I frame. Inception V3 0 confuses 4 of the 5 I frames. In all the misclassified frames the red areas are not focused on the metacarpal head joint.

In Table 5.4, VGG16 0 correctly classifies the NI frames. In fact, no significant activations are present in the classified frames. Inception V3 0 misidentifies 4 NI US scans, while focusing on the osteochondral interface. ResNet50 0 confuses the 4° and 5° NI frames.

Table 5.3: *In this table is represented the grad-CAM results for I frames. The green background stays for the correct classified frame; while the red background stays for the misclassified frame.*

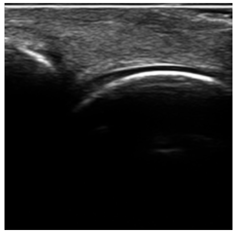
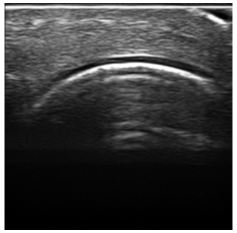
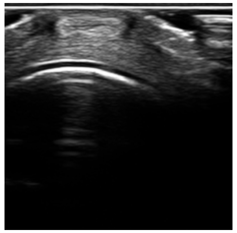
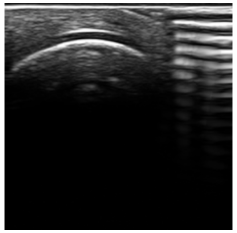
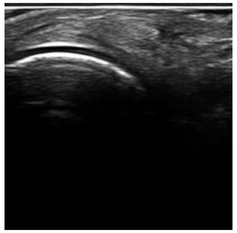
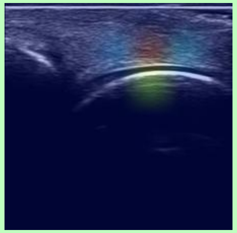
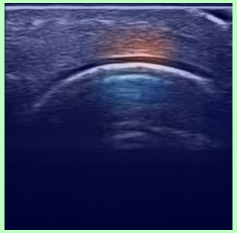
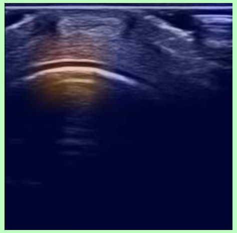
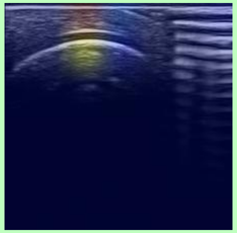
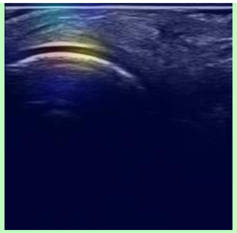
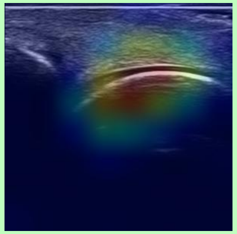
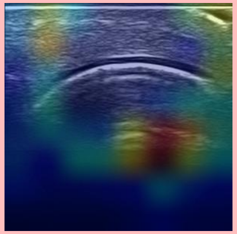
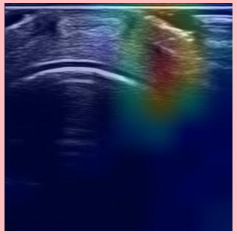
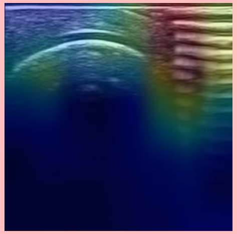
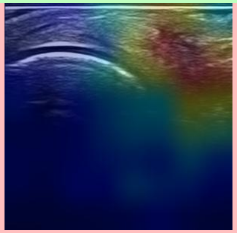
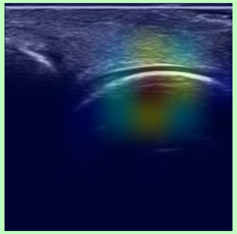
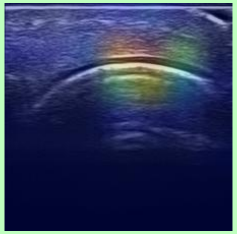
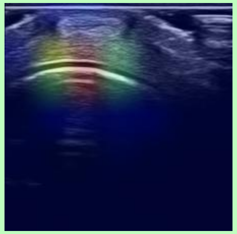
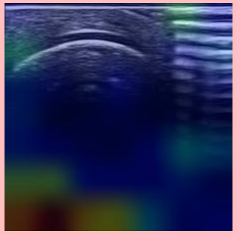
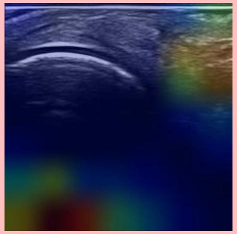
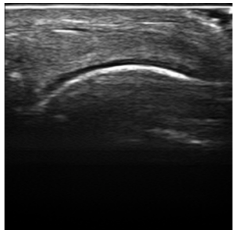
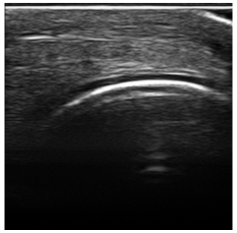
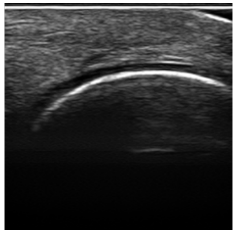
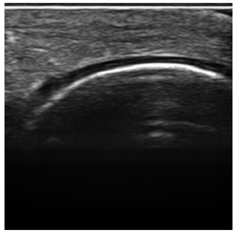
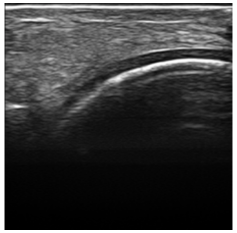
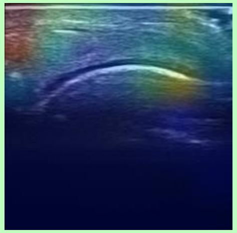
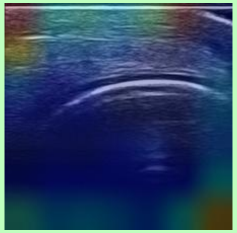
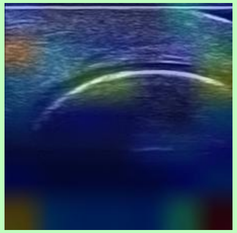
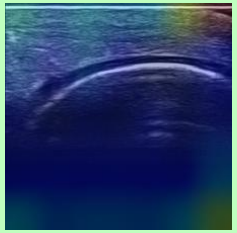
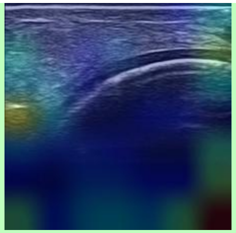
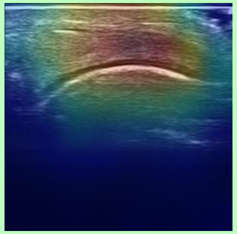
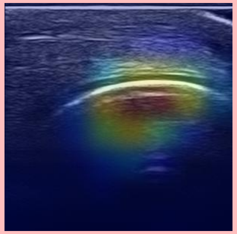
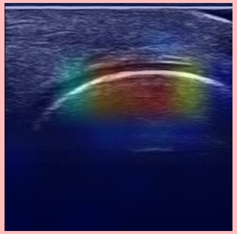
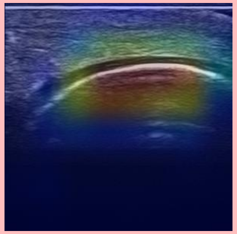
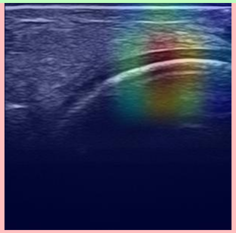
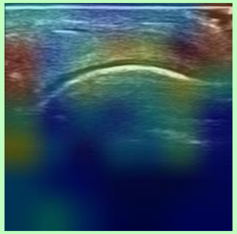
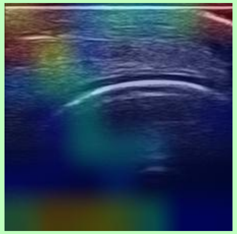
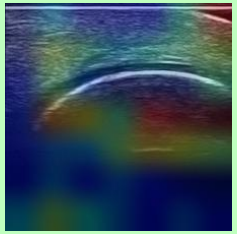
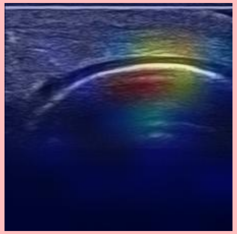
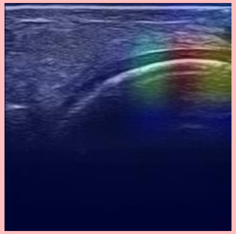
	1° I	**2° I**	**3° I**	**4° I**	**5° I**
					
VGG16 0					
Inception V3 0					
ResNet50 0					

Table 5.4: *In this table is represented the grad-CAM results for NI frames. The green background stays for the correct classified frame; while the red background stays for the misclassified frame.*

	1° NI	2° NI	3° NI	4° NI	5° NI
					
VGG16 0	🟩 	🟩 	🟩 	🟩 	🟩 
Incpetion V3 0	🟩 	🟥 	🟥 	🟥 	🟥 
ResNet50 0	🟩 	🟩 	🟩 	🟥 	🟥 

Chapter 6

DISCUSSION

Synopsis In this Chapter, the results presented in Chapter 5 are discussed. At the end of the Chapter, Section 6.1 reports the main limitations of this work and Section 6.2 illustrates the impact of the thesis.

In this work, a strategy for informative frame selection in metacarpal head US frames has been presented and evaluated. Instead of exploiting handcrafted features, as usually performed in the literature, this thesis investigated and assessed the performance of learned features extracted by means of six different transfer learning approaches on three CNNs, pre-trained on natural images, from 1945 frames of 40 different healthy patients. The classification was performed for two classes: informative and not informative. A leave 4-subjects out cross validation was repeated 10 times to perform a robust evaluation of classification performances.

The proposed approach resulted to be a valuable and robust tool to automatically identify informative frames in metacarpal heads US scanning, as shown by the high performances for I frames in terms of precision, recall, F1-score and AUC.

In the automatic frame selection literature, studies involving DL models, embrace the transfer learning approach as a method to overcome the lack of massive amount of clinical data as in this thesis. However in the literature [31, 45, 46] the proposed transfer learning approaches appeal to freeze the initial convolutional layers and to fine-tune the last ones as in [31] or to freeze the convolutional layers and to fine-tune the classifier layers as in [45, 46]. In this thesis, both of the previous approaches together with other 4 different transfer learning techniques have been compared for each selected CNN.

The aim of adopting six different transfer learning approaches is to identify the general to specific feature extraction transition with the purpose of increasing the performances by fine-tuning the specific features extraction.

The final surprising result is that for each selected CNN the best performing configuration is the one that involves the complete fine-tuning of the architecture. In fact, for the VGG16 the best of the six transfer learned configurations is the VGG16 0; for the Inception V3 is the Inception V3 0 and for the ResNet50 is the ResNet50 0.

Initializing the weights of the overall architecture with the transferred features and varying them during the training, lead to the extraction of learned features more suitable for the new classification task. Fine-tuning the initial convolutional layers, pointed to the extraction of general features, produce a boost in the performances of the new classification task, since they have been suited during the training to classify I and NI frames.

Thanks to the complete transfer learning approach the three CNNs VGG16 0, Inception V3 0 and ResNet50 0 reach an AUC higher than 0.99; leading to an almost perfect classification of I and NI frames. Among the three, the highest AUC is reached by the VGG16 0. The reason for this could be seen in the relatively simple (i.e. serial, without multiscale analysis or skip connections) architecture and small depth (16 layers) of the VGG16 architecture. Since the dataset does not have much variability the less depth of the VGG16 leads to the extraction of more generalizable features, resulting in a more successful transfer learning.

Looking at the behaviour of the three CNN inside the 10 leave 4-subjects out cross validation models (Figure 5.2) is possible to observe that VGG16 0 has the most stable way of working, respect to the ResNet50 0 and to the Incpetion V3 0. This is extrapolated by the narrower boxes that characterize the performances of the VGG16 0 for the I and NI frames. The outliers present in Figure 5.2 are performances that lie at an abnormal distance from all the other performances values. In particular, since the outliers are below the median and the mean value, they are synonymous of vulnerability.

It means that among the ten evaluation dataset there are some (depending on the number of outliers) in which the correct classification of I and NI failed more.

Considering the outliers of the VGG16 0, they never slump below the 0.9, meaning that performances can remain very high even with new dataset.

In Figure 5.2a, the NI precision of the Inception V3 0 falls below the 0.8 constituting a lack robustness of the CNN.

In order to evaluate the transfer learning approach, the training from scratch of the VGG16 has been performed. Comparing the AUC values of the mean ROC of the VGG16 0 and VGG16 from scratch, is possible to understand the efficiency of the adopted fine-tuning strategy.

The VGG16 0 outperforms the VGG16 from scratch in terms of class specific precision, recall, F1-score and in terms of discriminative power. This result is relevant and crucial in the medical field since collecting such a large amount of labeled images as in ImageNet is challenging, expensive and time consuming [59]. Transfer learning offers the possibility to overcome the data dependence problem and more over it helps in boosting the performances. Similar results are also found in [45] and in [46].

Considering Table 5.3, ResNet50 0 and Inception V3 0 misclassify more frames respect to the VGG16 0.

Looking at the behaviour of the VGG16 0 for the classification of I frames, it's astonishing to see how, despite the presence of US artifacts and not perfectly centered joint images, the VGG16 0 always focuses on the chondrosynovial interfaces. The red areas, in fact, always lie above this hyline cartilage edge. The VGG16 0 shows a perfect comprehension of the classification problem and of the informative predictor.

Considering the ResNet50 0 and the Inception V3 0 in the management of the I frames selection, they exhibit a lack of robustness in classifying I frames with US artifacts. In fact, in the 5° frame, instead of focusing on the metacarpal head joint, they center the attention on some US noise artifacts. A similar behaviour occurs in the classification of the 4° I frame. The frame presents a reverberation artifact on the right side of the sonogram and while the VGG16 0 centers the attention on the metacarpal head joint, the Inception V3 and ResNet50 focus respectively on the artifact and on the border of the image, misclassifing the frame. In the classification of 2° and 3° frame the Inception V3 0 draws the attention on other US artifacts present in the images. Accordingly to this results, Inception V3 0 and ResNet 50 lack of robustness because they are more vulnerable to US image artifacts.

In Table 5.4 ResNet50 0 and Inception V3 0 misclassify border line images (4°, 5°) in which the chondrosynovial interface is partially detectable or where speckle artifacts emulate the chondrosynovial interface. In the same frames the VGG16 0 does not show remarkable activations on the metacarpal joint, due to the lack of a significant presence of the chondrosynovial interface.

All of this considerations can be made thanks to grad-CAM. Grad-CAM allows to interpret why the CNN made a specific decision.

Newer regulations like the European General Data Protection Regulation (GDPR) are making harder the use of black-box models in all businesses including healthcare because interpretability of the decisions is now a requirement [60]. An AI system to complement medical professionals should have a certain amount of explainability and allow the human expert to retrace the decisions and use their judgment. In this thesis work the use of grad-CAM helps in building a more transparent algorithm that appropriately enhances trust of medical professionals in the DL model. By using grad-CAM the medical professional has the possibility to understand how and why a machine decision has been made.

Another use of Grad-CAM is identifying and reducing bias in the training dataset. Grad-CAM visualizations of the model predictions (Table 5.3 Table 5.4) revealed that the Inception V3 0 and ResNet 50 0 confuse US artifacts with the hyperechoic interfaces. This bias can be reduced by introducing in the dataset more US images with artifacts. Indeed, grad-CAM can help detect and remove biases in dataset, which is important for a better generalization of the model.

6.1 Limitations

A limit of the proposed work could be seen in the 'gold standard' image definition. The dataset was labeled by one expert sonographer and this may imply the risk of a systematic bias. Furthermore , the inter-reader realiability was not tested.

US images used in this study were obtained with the same US system machine, so the impact of using different manufactures' US system was not tested.

Some improvements (see Chapter 7) to the proposed work could be interesting to be further studied and potentially highly beneficial for the computer-assisted community.

6.2 Impact

US is an highly operator-dependent technique and sonographer skills and experience may affect both acquisition and interpretation process. Several international initiatives were undertaken to ensure the standardization of US assessment and to increase its reproducibility in rheumatological settings [30, 61]. The correct acquisition of an US image is the essential step to ensure an accurate and reliable assessment of the patient's conditions.

In this prospective the methodology proposed in this thesis is expected to impact positively on detection of musculoskeletal disease, allowing the clinicians to perform a more accurate US screening. The automatic frame selection system leads to a more objective assessment of image quality, increasing the US reproducibility and saving sonographers time. Furthermore, it may speed up the training process of beginner sonographers.

This strategy is also expected to impact positively on other body districts (i.e. hip, knees, ankle).

Although this thesis is the first attempt to exploit DL for informative-frame selection in US screening, the high performance achieved encourages for methodology translation into the actual clinical practice. The informative-frames selection is critical, as it could affect the output the computer-assisted support system and thus the judgment of the diagnosis. By supporting the informative frame selection process there is the possibility to enhance the accuracy of US joint scanning through a reduction in inter and intra-operator variability, as well as providing additional predictive information that may be too subtle to be detected by the human eyes.

Chapter 7

CONCLUSION AND FUTURE WORK

This thesis addresses the problem of informative-frame selection in US metacarpal head joint scanning. The proposed approach exploited the advantages of DL models, using six different transfer learning for each selected CNN (VGG16, Inception V3 and ResNet50).

The method was validated in ten different models achieving impressive results. For the best performing CNN (i.e. VGG16 0) a value of 0.99 was reached for the precision, recall and F1-score of I frames, beating the state-of-art performances.

The work does not just offer the possibility of an almost perfect classification of informative and not informative frames, but it also provides a visual explanation of the DL decision. Thanks to this human experts can understand and retrace the machine decision leading to a synergic work between the humans and AI that empowers the accuracy of predictions.

Being the dimension of the dataset the main limitation of this work, enlarging the dataset by exploiting Generative Adversarial Networks could enable better fine-tuning of the proposed system. This should attenuate misclassification problems allowing the proposed methodology to be easily and successfully integrated as pre-processing step for the computer-assisted diagnosis systems.

As next step pathological images could also be included in the dataset to encode further challenges. It could be also interesting enlarging the dataset by including different anatomical districts as knee, ankle, hip, wrist, shoulder. A further action could be the introduction of more classes like informative, not informative and partially informative.

As future work it could be also exciting a real-time integration of the algorithm during the US scanning so that the clinician has a instantaneous feedback while performing the US exam. With a real-time application of the automatic I frame selection young residents have the chance to learn faster and to reach an operational US autonomy in a shorter period of time.

To conclude the present work represents an effective tool to support the clinician in US scanning and to lead to more objective evaluation of the patient's condition reducing the operator-dependency factor that characterizes the US imaging technique.

List of Abbreviations

RA: Rheumatoid Arthritis

ACR: America College of Rheumatology

EULAR: European League Against Rheumatism

RF: Rheumatoid Factor

ACPA: Anti-Citrullinated Protein Antibody

CRP: C-Reactive Protein

ESR: Erythrocyte sedimentation rate

CR: Conventional radiography

US: Ultrasound

I: Informative

NI: Not informative

AI: Artificial intelligence

ML: Machine learning

DL : Deep learning

CNN: Convolutional Neural Network

GLCM: Grey Level co-occurance matrix

HSV: Hue-saturation-value

WCE: Wireless capsule endoscopy

SVM: Support Vector Machine

RGB: Red blue green

Grad-CAM: Gradient-weighted Class Activation Mapping

ReLU: Rectified Linear Unit

VGG: Visual Geometric Group

ILSVRC: ImageNet Large Scale Visual Recognition Challenge

SGD: Stochastic Gradient Descend

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

AUC: Area under the curve

ROC: Receiver Operating Characteristics

GDPR: General Data Protection Regulation

List of Figures

List of Tables

Bibliography

[1] M. Kourilovitch, C. Galarza-Maldonado, and E. Ortiz-Prado, “Diagnosis and classification of rheumatoid arthritis,” *Journal of autoimmunity*, vol. 48, pp. 26–30, 2014.

[2] F. Wolfe, “The natural history of rheumatoid arthritis.,” *The Journal of rheumatology. Supplement*, vol. 44, p. 13, 1996.

[3] Y.-N. Bi, C.-H. Xiao, C. Pan, X.-F. Zhao, Y.-Y. Cao, Y. Yi, and F.-F. Zuo, “The correlation study on syndrome differentiation of rheumatoid arthritis and joint high frequency ultrasound performance,” *Zhongguo Zhong xi yi jie he za zhi Zhongguo Zhongxiyi jiehe zazhi= Chinese journal of integrated traditional and Western medicine*, vol. 35, no. 1, pp. 19–24, 2015.

[4] J. Adams, J. Burridge, M. Mullee, A. Hammond, and C. Cooper, “Correlation between upper limb functional ability and structural hand impairment in an early rheumatoid population,” *Clinical Rehabilitation*, vol. 18, no. 4, pp. 405–413, 2004.

[5] S. Laivoranta-Nyman, R. Luukkainen, M. Hakala, P. Hannonen, T. Möttönen, U. Yli-Kerttula, J. Ilonen, and A. Toivanen, “Differences between female and male patients with familial rheumatoid arthritis,” *Annals of the rheumatic diseases*, vol. 60, no. 4, pp. 413–415, 2001.

[6] S. Viatte, D. Plant, and S. Raychaudhuri, “Genetics and epigenetics of rheumatoid arthritis,” *Nature Reviews Rheumatology*, vol. 9, no. 3, p. 141, 2013.

[7] Y. P. Goekoop-Ruiterman, J. K. de Vries-Bouwstra, C. F. Allaart, D. van Zeben, P. J. Kerstens, J. M. W. Hazes, A. H. Zwinderman, A. J. Peeters, J. M. de Jonge-Bok, C. Mallée, *et al.*, "Comparison of treatment strategies in early rheumatoid arthritis: a randomized trial," *Annals of internal medicine*, vol. 146, no. 6, pp. 406–415, 2007.

[8] D. Aletaha, T. Neogi, A. J. Silman, J. Funovits, D. T. Felson, C. O. Bingham III, N. S. Birnbaum, G. R. Burmester, V. P. Bykerk, M. D. Cohen, *et al.*, "2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative," *Arthritis & rheumatism*, vol. 62, no. 9, pp. 2569–2581, 2010.

[9] B. Heidari, Z. Lotfi, R. Ali Firouzjahi, and P. Heidari, "Comparing the diagnostic values of anti-cyclic citrullinated peptide antibodies and rheumatoid factor for rheumatoid arthritis," *Research in medicine*, vol. 33, no. 3, pp. 156–161, 2010.

[10] P. Emery, I. McInnes, R. Van Vollenhoven, and M. Kraan, "Clinical identification and treatment of a rapidly progressing disease state in patients with rheumatoid arthritis," *Rheumatology*, vol. 47, no. 4, pp. 392–398, 2008.

[11] B. Heidari, "Rheumatoid arthritis: Early diagnosis and treatment outcomes," *Caspian journal of internal medicine*, vol. 2, no. 1, p. 161, 2011.

[12] J. Yue, D. Wu, and L.-S. Tam, "The role of imaging in early diagnosis and prevention of joint damage in inflammatory arthritis," *Expert review of clinical immunology*, vol. 14, no. 6, pp. 499–511, 2018.

[13] I. Castrejón, L. McCollum, M. D. Tanriover, and T. Pincus, "Importance of patient history and physical examination in rheumatoid arthritis compared to other chronic diseases: results of a physician survey," *Arthritis care & research*, vol. 64, no. 8, pp. 1250–1255, 2012.

[14] A. N. Colebatch, C. J. Edwards, M. Østergaard, D. Van Der Heijde, P. V. Balint, M.-A. D'Agostino, K. Forslind, W. Grassi, E. A. Haavardsholm, G. Haugeberg, *et al.*, "Eular recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis," *Annals of the rheumatic diseases*, vol. 72, no. 6, pp. 804–814, 2013.

[15] A. Saraux, J. M. Berthelot, G. Chalès, C. Le Henaff, J. B. Thorel, S. Hoang, I. Valls, V. Devauchelle, A. Martin, D. Baron, *et al.*, "Ability of the american college of rheumatology 1987 criteria to predict rheumatoid arthritis in patients with early arthritis and classification of these patients two years later," *Arthritis & Rheumatism*, vol. 44, no. 11, pp. 2485–2491, 2001.

[16] M. Østergaard, "Can imaging be used for inflammatory arthritis screening?," in *Seminars in musculoskeletal radiology*, vol. 16, pp. 401–409, Thieme Medical Publishers, 2012.

[17] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: a review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[18] W. M. Hetta, S. M. Sharara, and G. A. Gouda, "Role of magnetic resonance imaging and ultrasonography in diagnosis and followup rheumatoid arthritis in hand and wrist joints," *The Egyptian Journal of Radiology and Nuclear Medicine*, vol. 49, no. 4, pp. 1043–1051, 2018.

[19] M. Boylan, "Should ultrasound be used routinely in the diagnosis of rheumatoid arthritis?," *Irish Journal of Medical Science (1971-)*, vol. 189, no. 2, pp. 735–748, 2020.

[20] M. Jain and J. Samuels, "Musculoskeletal ultrasound in the diagnosis of rheumatic disease," *Bulletin of the NYU Hospital for Joint Diseases*, vol. 68, no. 3, p. 183, 2010.

[21] H. Xu, Y. Zhang, H. Zhang, C. Wang, and P. Mao, "Comparison of the clinical effectiveness of us grading scoring system vs mri in the diagnosis of early rheumatoid arthritis (ra)," *Journal of orthopaedic surgery and research*, vol. 12, no. 1, p. 152, 2017.

[22] A. Kumar, "How to investigate new-onset polyarthritis," *Best Practice & Research Clinical Rheumatology*, vol. 28, no. 6, pp. 844–859, 2014.

[23] F. Porta, G. Radunovic, V. Vlad, M. C. Micu, R. Nestorova, T. Petranova, and A. Iagnocco, "The role of doppler ultrasound in rheumatic diseases," *Rheumatology*, vol. 51, no. 6, pp. 976–982, 2012.

[24] A.-B. Aga, H. B. Hammer, I. C. Olsen, T. Uhlig, T. K. Kvien, D. van der Heijde, H. Fremstad, T. M. Madland, Å. S. Lexberg, H. Haukeland, *et al.*, "First step in the development of an ultrasound joint inflammation score for rheumatoid arthritis using a

data-driven approach," *Annals of the rheumatic diseases*, vol. 75, no. 8, pp. 1444–1451, 2016.

[25] J. S. Smolen, F. C. Breedveld, G. R. Burmester, V. Bykerk, M. Dougados, P. Emery, T. K. Kvien, M. V. Navarro-Compán, S. Oliver, M. Schoels, *et al.*, "Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force," *Annals of the rheumatic diseases*, vol. 75, no. 1, pp. 3–15, 2016.

[26] D. Nakagomi, K. Ikeda, A. Okubo, T. Iwamoto, Y. Sanayama, K. Takahashi, M. Yamagata, H. Takatori, K. Suzuki, K. Takabayashi, *et al.*, "Ultrasound can improve the accuracy of the 2010 american college of rheumatology/european league against rheumatism classification criteria for rheumatoid arthritis to predict the requirement for methotrexate treatment," *Arthritis & Rheumatism*, vol. 65, no. 4, pp. 890–898, 2013.

[27] A. Filer, P. De Pablo, G. Allen, P. Nightingale, A. Jordan, P. Jobanputra, S. Bowman, C. D. Buckley, and K. Raza, "Utility of ultrasound joint counts in the prediction of rheumatoid arthritis in patients with very early synovitis," *Annals of the rheumatic diseases*, vol. 70, no. 3, pp. 500–507, 2011.

[28] M. A. D'Agostino, L. Terslev, R. Wakefield, M. Østergaard, P. Balint, E. Naredo, A. Iagnocco, M. Backhaus, W. Grassi, and P. Emery, "Novel algorithms for the pragmatic use of ultrasound in the management of patients with rheumatoid arthritis: from diagnosis to remission," *Annals of the rheumatic diseases*, vol. 75, no. 11, pp. 1902–1908, 2016.

[29] F. Salaffi, M. Gutierrez, and M. Carotti, "Ultrasound versus conventional radiography in the assessment of bone erosions in rheumatoid arthritis," *Clin Exp Rheumatol*, vol. 32, no. 1 Suppl 80, pp. S85–S90, 2014.

[30] P. Mandl, P. Studenic, E. Filippucci, A. Bachta, M. Backhaus, D. Bong, G. A. Bruyn, P. Collado, N. Damjanov, C. Dejaco, *et al.*, "Development of semiquantitative ultrasound scoring system to assess cartilage in rheumatoid arthritis," *Rheumatology*, vol. 58, no. 10, pp. 1802–1811, 2019.

[31] M. C. Fiorentino, S. Moccia, E. Cipolletta, E. Filippucci, and E. Frontoni, "A learning approach for informative-frame selection in us rheumatology images," in *International Conference on image analysis and processing*, pp. 228–236, Springer, 2019.

[32] Q. Huang, F. Zhang, and X. Li, “Machine learning in ultrasound computer-aided diagnostic systems: a survey,” *BioMed research international*, vol. 2018, 2018.

[33] L. Nanni, S. Ghidoni, and S. Brahnam, “Handcrafted vs. non-handcrafted features for computer vision classification,” *Pattern Recognition*, vol. 71, pp. 158–172, 2017.

[34] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, “How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?,” *arXiv preprint arXiv:1511.06348*, 2015.

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[36] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. C. de Groen, “Informative frame classification for endoscopy video,” *Medical Image Analysis*, vol. 11, no. 2, pp. 110–127, 2007.

[37] C. Ballesteros, M. Trujillo, and C. Mazo, “Automatic classification of non-informative frames in colonoscopy videos,” 2015.

[38] M. A. Armin, G. Chetty, F. Jurgen, H. De Visser, C. Dumas, A. Fazlollahi, F. Grimpen, and O. Salvado, “Uninformative frame detection in colonoscopy through motion, edge and color features,” in *Computer-Assisted and Robotic Endoscopy*, pp. 153–162, Springer, 2015.

[39] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori, “Automatic detection of informative frames from wireless capsule endoscopy images,” *Medical Image Analysis*, vol. 14, no. 3, pp. 449–470, 2010.

[40] I. Mehmood, M. Sajjad, and S. W. Baik, “Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure,” *Journal of medical systems*, vol. 38, no. 9, p. 109, 2014.

[41] O. H. Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, M. Alizadeh, and H. A. Soleimani, “Informative and uninformative regions detection in wce frames,” *Journal of Advanced Computing*, vol. 3, no. 1, pp. 12–34, 2014.

[42] A. Ishijima, R. A. Schwarz, D. Shin, S. Mondrik, N. Vigneswaran, A. M. Gillenwater, S. Anandasabapathy, and R. Richards-Kortum, “Automated frame selection process for

high-resolution microendoscopy," *Journal of biomedical optics*, vol. 20, no. 4, p. 046014, 2015.

[43] A. Perperidis, A. Akram, Y. Altmann, P. McCool, J. Westerfeld, D. Wilson, K. Dhaliwal, and S. McLaughlin, "Automated detection of uninformative frames in pulmonary optical endomicroscopy," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 1, pp. 87–98, 2016.

[44] S. Moccia, G. O. Vanone, E. De Momi, A. Laborai, L. Guastini, G. Peretti, and L. S. Mattos, "Learning-based classification of informative laryngoscopic frames," *Computer methods and programs in biomedicine*, vol. 158, pp. 21–30, 2018.

[45] A. R. Islam, A. Alammari, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Non-informative frame classification in colonoscopy videos using cnns," in *Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing*, pp. 53–60, 2018.

[46] I. Patrini, M. Ruperti, S. Moccia, L. S. Mattos, E. Frontoni, and E. De Momi, "Transfer learning for informative-frame selection in laryngoscopic videos through learned features," *Medical & Biological Engineering & Computing*, pp. 1–14, 2020.

[47] M. R. Zare, D. O. Alebiosu, and S. L. Lee, "Comparison of handcrafted features and deep learning in classification of medical x-ray images," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 1–5, IEEE, 2018.

[48] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine & Biology*, vol. 62, no. 19, p. 7714, 2017.

[49] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, 2016.

[50] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[52] D. CireşAn, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural networks*, vol. 32, pp. 333–338, 2012.

[53] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*, pp. 270–279, Springer, 2018.

[54] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[56] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[57] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[58] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[59] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[60] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, 2017.

[61] I. Möller, I. Janta, M. Backhaus, S. Ohrndorf, D. A. Bong, C. Martinoli, E. Filippucci, L. M. Sconfienza, L. Terslev, N. Damjanov, *et al.*, "The 2017 eular standardised procedures for ultrasound imaging in rheumatology," *Annals of the rheumatic diseases*, vol. 76, no. 12, pp. 1974–1979, 2017.