UNIVERSITÀ DEGLI STUDI DI MILANO

Mitigating Data Scarcity Challenges in Medical Imaging Analysis:

Advanced Learning Approaches with Emphasis on Hemophilic UltraSound images

Marco Colussi
PhD School in Computer Science
XXXVII Cycle

Computer Science Department
"Giovanni Degli Antoni"

Advisor: **Prof. Sergio Mascetti**
Co-advisor: **Prof. Claudio Bettini**

PhD Coordinator: **Prof. Roberto Sassi**

A.A. 2023/2024

Abstract

Medical imaging plays a crucial role in hemophilia research and clinical practice, particularly in assessing joint health and bleeding events. Ultrasound (US) imaging is a fundamental tool in the diagnostic process and is currently used to identify when the joint recess is filled with synovial fluid or blood, a condition known as “recess distention” that, if filled with blood, can potentially lead to pathologies and permanent joint damage. In this context, deep learning (DL) techniques can support image acquisition (possibly at the point-of-care) and enhance the capabilities of computer-aided diagnosis (CAD) systems.

However, the lack of labeled training data makes the effective utilization of DL techniques in the medical domain impractical, leading to suboptimal performance in various imaging tasks, such as classification, detection, and segmentation. This thesis investigates the application of advanced DL methods to overcome this challenge and enhance the analysis of medical images in the context of data scarcity, with a particular focus on ultrasound images in hemophilia research. Specifically, this thesis addresses three main challenges: a limited number of total samples, class imbalance, and the adaptation of trained models to different domains (such as knee to elbow transfer).

To address the problem of the limited number of total samples, this research investigates the adoption of transfer learning and proposes a new multi-task model to effectively utilize limited labeled data and improve model generalization. Concerning the issue of imbalanced data, the thesis explores anomaly detection techniques that can be trained on normal samples only. However, as demonstrated experimentally, classic unsupervised anomaly detection methods fail in this domain due to the intrinsic variability of musculoskeletal ultrasound images. Therefore, the thesis introduces a new weakly supervised anomaly detection framework that enhances classification and segmentation performance, requiring only the recess location as a weak annotation. To address the third issue, we investigate two domain adaptation frameworks to adapt a model trained on knee images to also identify the distension on elbow images. We first explore test-time adaptation techniques and then introduce a new contrastive feature test-time training approach.

By developing and integrating these DL techniques into an existing CAD system, this thesis aims to provide insights into effectively leveraging limited labeled data in medical imaging research, thereby advancing the understanding and management of rare and complex medical conditions.

Author's Publications

This thesis is founded upon the publications authored throughout my three-years of doctoral degree.

Journals

- Colussi, M., Civitarese, G., Ahmetovic, D., Bettini, C., Gualtierotti, R., Peyvandi, F., & Mascetti, S. (2023). *Ultrasound detection of subquadricipital recess distension.* Intelligent Systems with Applications, 17, 200183.[1]

- Campana, M. G., Colussi, M., Delmastro, F., Mascetti, S., & Pagani, E. (2024). *A Transfer Learning and Explainable Solution to Detect mpox from Smartphones images.* Pervasive and Mobile Computing, 98, 101874.[2]

- Gualtierotti, R., Giachi, A., Suffritti, C., Bedogni, L., Franco, F., Poggi, F., Mascetti, S., Colussi, M., Ahmetovic, D., Begnozzi, V., Boccalandro, E. A., Solimeno, L. P. & Peyvandi, F. (2024). *Optimizing long-term joint health in the treatment of hemophilia.* Expert Review of Hematology, 44.[3]

International conferences

- Colussi, M., Mascetti, S., Dolz, J. and Desrosiers, C. (2024). *ReC-TTT: Contrastive Feature Reconstruction for Test-Time Training.* (Accepted for publication @ WACV2025).

International workshops

- Colussi, M., Mascetti, S., Ahmetovic, D., Civitarese, G., Cacciatori, M., Peyvandi, F., ... & Bettini, C. (2023, October). *GAJA-Guided self-Acquisition of Joint ultrAsound images.* In International Workshop on Advances in Simplifying Medical Ultrasound held in conjunction with the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 132-141). Cham: Springer Nature Switzerland.[4]

- Ahmetovic D., Angileri A., Arcudi S., Bettini C., Civitarese G.,Colussi M., ... & Truma A. (2024, June). *Insights on the development of PRACTICE,a research-oriented healthcare platform.* In 2024 IEEE International Conference on Smart Computing (SMARTCOMP).[5]

- Colussi, M., Mascetti, S., Ahmetovic, D., Civitarese, G., Cacciatori, M., Peyvandi, F., ... & Bettini, C. (2024, October). *LoRIS: Weakly-supervised Anomaly Detection for Ultrasound Images.* In International Workshop on Advances in Simplifying Medical Ultrasound held in conjunction with the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Cham: Springer Nature Switzerland.[6]

Acknowledgments

I would like to begin by expressing my gratitude to my supervisor, Sergio Mascetti, for his constant guidance, always pointing me in the right direction. His help extended far beyond the role of supervisor, and I am deeply thankful for the assistance he provided. I am also thankful to my co-supervisor, Claudio Bettini, for the constructive discussions we shared throughout these years, and to Roberta Gualtierotti for providing all the medical expertise required for the project. My sincere thanks go to all the members of the EwLab, Michele, Gabriele G., Dragan, Gabriele C., and Matteo, with whom I shared many joyful moments and who patiently listened to my complaints and frustrations, always offering their support.

I would like to extend my gratitude to Christian Desrosiers and Jose Dolz for warmly welcoming me to their laboratory and dedicating their time and knowledge during my stay in Montreal. Additionally, I am grateful to have met wonderful people at the LIVIA Lab, who I have shared ideas and happy moments with.

This version of the thesis would not have been possible without the insightful suggestions and excellent work of Christian Desrosiers, Alberto Gomez, and Stergios Christodoulidis, who served as thesis reviewers and whose contributions greatly enhanced its structure and quality.

Thank you to Carlo, Giro, Sullo, Alice, Umbi, Manuel, and all the friends who made everything easier. Your encouragement and friendship have meant the world to me.

Finally, I would like to express my heartfelt thanks to my family,

mom, dad, and Giuggi, for always being by my side and providing me with the support and tools needed to reach this point. And Silvia for accompanying me on this journey and the ones to come, always offering the right words, and for being a constant source of strength. P.S. A special thanks as well for contributing to half of the graphs in this thesis.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Context

Hemophilia [1] is a hereditary blood coagulation disorder that results in an increased risk of spontaneous bleeding or due to trauma, which worsens with the severity of the disease. Bleedings can also frequently occur inside joints (mostly ankles, knees, and elbows) and muscles, which together account for around 80% of the bleeding events in patients with hemophilia [8, 9]. Joint bleeding causes the distension of the affected joint recess and, if not promptly treated with coagulation factor (Factor VIII or IX), can result in permanent damage such as synovial hyperplasia, osteochondral damage, and hemophilic arthropathy [10]. Thus, it is essential to promptly recognize joint recess distension.

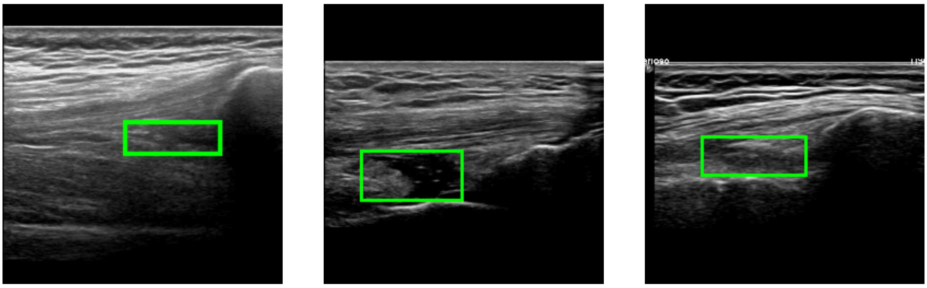
Physical examination may not be sufficient to diagnose joint recess distention, as in the early stage it can be asymptomatic [11].

[1] A glossary of all medical terms used is provided in the Appendix E.

Magnetic Resonance Imagining (MRI) is generally considered the gold standard tool for precise evaluation of joints, but is not practical for regular follow-up of patients with hemophilia due to high costs, limited availability, and long examination times [11]. An alternative solution is ultrasound (US) imaging [12] that, contrary to MRI, is low cost, has a short examination time, and is widely accessible [13]. An example of a standardized protocol for US imaging in hemophilia is the *Hemophilia Early Arthropathy Detection with UltraSound* (HEAD-US), designed to guide the practitioner in acquiring relevant US images and interpreting them for the diagnosis of joint recess distension in the joints 6 most commonly affected [14].

A joint recess can be *Distended* for three main reasons: if it is filled with synovial liquid, if it is filled with blood (a condition known as *blood effusion*), and if its membrane is thicker due to an inflammation known as *synovitis*. Figure 1.1 shows three examples of the knee longitudinal subquadricipital recess (SQR) scan, one of the possible US views acquired using the HEAD-US protocol (a list is available in Appendix D). In Figure 1.1a the SQR is the dark area shown in the green box. In this case, the SQR is thin and hence it is *Non-distended*. Vice versa, in Figure 1.1b the SQR is much thicker, indicating that it is *Distended*. While Figure 1.1a and 1.1b show two characteristic examples with stark differences, there are borderline cases where the SQR appears slightly enlarged but it is *Non-distended* (see Figure 1.1c) or it is very slightly *Distended*. The SQR in knee ultrasound is not always clearly visible. Therefore, its approximate position can be inferred from the location of the three characterizing elements (*i.e.*, patella, femur, and tendons): the recess is positioned below the tendons, above the rightmost end of the femur, and on the patella bottom left. To determine the exact position of the recess, the practitioner observes the anechogenic area present in the region. The recess appears as a dark area surrounded by a lighter membrane. A more detailed description is provided in Section 2.1

To determine whether the SQR is *Distended*, practitioners qualitatively establish whether it is swollen, a sign that it is filled with liquid, or that its membrane is thickened. Instead, a *Non-distended* recess commonly appears as a thin line. We highlight that the use of subjective assessment of imaging data as ground truth is a common practice in clinical evaluation [15].



(a) *Non-distended* SQR

(b) *Distended* SQR

(c) Borderline *Non-distended* SQR

Figure 1.1: Examples of longitudinal SQR scans

1.2 Motivation

One unmet need in the hemophilia management is that patients are required to visit a medical facility in case of suspect blood effusion. However, many patients are not always able to make these visits, and in some cases they may take the coagulant factor solely on the basis of pain level, which can lead to overtreatment or undertreatment.

A possible solution to address this problem is to provide a tool to support patients in self-acquiring images at their point-of-care (POC).

Although such a solution would address an unmet need, it would also generate large amounts of images that would need to be analyzed by skilled practitioners to perform a diagnosis. However, the availability of such figures is limited. To address this issue, a Computer-Aided Diagnosis (CAD) system can be adopted to support the medical practitioner in the diagnosis process by distinguishing between

Distended or *Non-distended* recesses and speeding up the process hence ensuring timely interventions.

By addressing these objectives, we plan to significantly enhance the management of hemophilic patients, leading to better outcomes and more efficient use of healthcare resources. However, these goals are constrained by the limited amount of data available for training deep learning models. Addressing this challenge is the central focus of the thesis, which explores methods to overcome data scarcity and improve the performance of automated diagnostic systems.

1.3 Challenges

Despite the significant success of computer vision research in various medical imaging tasks, most approaches rely heavily on large curated datasets. These datasets are required to correctly train deep learning models to accurately perform vision tasks such as classification, detection, and segmentation. However, in the medical context, the acquisition of these data sets presents significant challenges. Sharing personal medical data raises privacy concerns, which complicates the collection of diverse datasets from multiple centers. Furthermore, issues related to data usage rights, such as the requirement for informed consent specifying the purpose and duration of data use, as well as licensing restrictions set by data custodians, add to the complexity. Furthermore, the acquisition of medical images is often restricted due to limited patient data, the rarity of the pathology, and the availability of high-quality imaging equipment.

The difficulty of obtaining detailed annotations, such as image class annotations, bounding boxes, and segmentation masks, poses an even more challenging scenario. These types of annotations require skilled practitioners to analyze each individual image and possibly to identify complex anatomical boundaries. This is a labor-intensive and time-consuming process due to the inherent complexity of the domain. Furthermore, some features of anatomical images may not be clearly visible or consistent across images of different patients, which, together with the high noise and variability of the images, adds more complexity to the annotation process. Consequently, the lack of largely annotated data poses a significant barrier to developing robust and generalizable models in these settings.

This thesis addresses three different challenges related to the problems described above: the limited number of total samples, the imbalance between classes, and the adaptation of trained models to different unlabeled domains.

Limited number of total sample. The novelty of certain medical conditions, such as the early stages of COVID-19 and MPOX, as well as the rarity of pathologies such as hemophilia, makes it difficult to collect large amounts of labeled data. This limitation restricts the ability of deep learning models to generalize effectively. This thesis addresses various challenges that arise when available data is extremely limited. First, transfer learning is employed to leverage knowledge from larger datasets. Then, multi-task learning is used to extract and apply knowledge across different tasks, further enhancing the model's performance.

Class imbalance. Beyond the rarity of the pathology, the collection of images with blood effusion requires patients to have active swelling, which is often not the case for two reasons. First, recent advancements in hemophilia treatment have reduced the frequency and intensity of swelling. Second, as mentioned before, patients may not always be able to reach a medical facility, making it difficult to capture images of this condition. This results in class imbalance, with *Non-distended* images dominating the dataset, leading to models that have low sensitivity, an issue of particular importance in the medical field. To solve this issue, we explored the anomaly detection framework, where the training of the model is performed using only *Non-distended* images and adopting the bounding box of the recess as weak supervision.

Adaptation to different domains. As described in detail in Section 2.1, hemophilia primarily affects the knee, which is more prone to impact and injury. However, it is also crucial to identify swelling in other similar joints, such as the elbow. Training a model to classify images of the elbow recess would typically require collecting and annotating a large set of images, since deep learning models are known to struggle in generalizing on new unseen domains. However, due to the similarity between the elbow and knee joints, we employed domain adaptation techniques to avoid retraining the model in a fully supervised manner and instead addressed the domain shift between knee and elbow recesses.

1.4 Contributions

This thesis describes the contributions of the work conducted during my three years at Everyware Lab, in collaboration with the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, and during the period spent at the LIVIA Lab of ETS Montreal. The work addresses four main challenges: data scarcity, data imbalance, domain-shift, and system integration.

1.4.1 Data scarcity

To tackle the previously described problem of having limited amounts of data available, we initially investigated the adoption of transfer learning (TL) to detect mpox from skin lesion images. In this recently spread pathology, publicly available datasets in the literature are limited and corrupted due to a web-scraping dataset generation approach. We initially collected and curated a small dataset of skin lesions, named Mpox Close Skin Images (MCSI). We applied TL to leverage the strengths of pre-trained networks, enhancing model performances. To do so, we fine-tuned and compared five state-of-the-art deep learning models. This was achieved with a comprehensive evaluation using a 10-fold cross-validation technique to ensure the robustness and generalizability of our models. Furthermore, we optimized the best-performing model for mobile device use, taking into account the typical memory limitations of mobile devices. This optimization enables all data processing and classification to be performed directly on the device without the need of a connection, not always available in rural areas where this pathology is more prevalent.

TL showed promising results on the MPOX dataset, but not satisfactory in the US musculoskeletal domain, prompting us to reformulate the problem to specifically focus on detecting and classifying the SQR and its distension. In this context, we collected and annotated an SQR knee US dataset with the collaboration of the Policlinico of Milano. To tackle this new problem, we proposed two different solutions: one based on a single-stage detection task, and the other utilizing a multi-task learning approach. We then evaluated and compared these proposed solutions using the newly collected dataset, providing a comprehensive analysis of their performance. This showed promising results in terms of pathology detection and accurate recess detection ability, outperforming the simpler TL approach in both tasks.

These contributions are reported in Chapter 3 and Chapter 4 and are based on the following pubblications:

- Campana, M. G., Colussi, M., Delmastro, F., Mascetti, S., & Pagani, E. (2024). *A Transfer Learning and Explainable Solution to Detect mpox from Smartphones images.* Pervasive and Mobile Computing, 98, 101874.

My contributions

- Collaboration in concept and methodology design.
- Collaboration in data collection and cleaning.
- Collaboration in method implementation.
- Collaboration in the design of the evaluation methods.
- Collaboration in experiments execution.
- Collaboration in results analysis and interpretation.

- Colussi, M., Civitarese, G., Ahmetovic, D., Bettini, C., Gualtierotti, R., Peyvandi, F., & Mascetti, S. (2023). *Ultrasound detection of subquadricipital recess distension.* Intelligent Systems with Applications, 17, 200183.

My contributions

- Collaboration in concept and methodology design.
- Collaboration in the design of data collection protocol and annotation tool.
- Method implementation.
- Collaboration in the design of the evaluation methods.
- Experiments execution.
- Collaboration in results analysis and interpretation.

1.4.2 Data imbalance

The approach proposed in Chapter 4, while demonstrating promising results, has limitations in accurately identifying *Distended* cases. This is likely due to the imbalance of the training data. Additionally, it only provides the bounding box of the recess, while its segmentation offers more insight and assists practitioners in better utilizing the suggestions provided.

To address these limitations, we propose LoRIS, the first weakly supervised anomaly detection and segmentation technique specifically designed for US knee images. LoRIS addresses the critical issue of class imbalance in medical imaging by leveraging *Non-distended* recess data during training to detect anomalies effectively. Unlike traditional fully supervised methods that require extensive labeled datasets, LoRIS can be trained using only images from a single class, making it a more feasible option for medical applications where annotated data is scarce. Additionally, LoRIS not only detects anomalies but also provides *Distended* recess segmentations requiring only the weak supervision of the bounding boxes.

Through comprehensive evaluation and ablation studies, we demonstrate that current state-of-the-art unsupervised anomaly detection methods are not effective in this domain. In contrast, LoRIS shows comparable performance to supervised solutions, thus validating its effectiveness. Its primary advantage over supervised techniques lies in its ability to operate without the need for exhaustive labeled data. Lastly, we present an automated method for computing the location prior, enabling a fully automated detection pipeline during inference. This advancement further improves the practicality of LoRIS, reducing the need for manual intervention.

This contribution is described in Chapter 5 and it's based on the following publication:

- Colussi, M., Mascetti, S., Ahmetovic, D., Civitarese, G., Cacciatori, M., Peyvandi, F., ... & Bettini, C. (2024, October). *LoRIS: Weakly-supervised Anomaly Detection for Ultrasound Images.* In International Workshop on Advances in Simplifying Medical Ultrasound. Cham: Springer Nature Switzerland.

My contributions

- Concept and methodology design.
- Method implementation.
- Collaboration in the design of the evaluation methods.
- Experiments execution.
- Collaboration in results analysis and interpretation.

1.4.3 Domain-shift

Since blood effusion can affect various joints, we employed a domain adaptation approach to address the domain shift between knee SQR images and elbow OLR images. Despite the anatomical differences between these joints, the recess and its potential distention exhibit significant similarities, making domain adaptation a suitable method to improve model performance across these different joint types.

To tackle this problem, we first propose ReC-TTT, the first test-time training (TTT) approach that leverages contrastive feature reconstruction as a self-supervised task. By incorporating contrastive learning, our method enables the model to effectively distinguish subtle differences in feature representations, correctly adapt the encoders to the new domain, and enhance its robustness to shifts. Furthermore, we added an ensemble learning strategy in which two classifiers are trained using different image augmentations to ensure consistent predictions. Comprehensive experiments on different types of datasets with different types of distribution shifts, supported by extensive ablation studies, demonstrate that our method outperforms recent test-time adaptation (TTA) and TTT techniques, achieving state-of-the-art performance with fewer parameters to tune and its more robust on smaller batches.

An evaluation conducted on a **Dataset** demonstrates the potential to adapt to new datasets without requiring re-annotation also in the US context. In particular, it shows that by using test-time training (TTT), we can successfully adapt from the SQR dataset to the OLR dataset, achieving good adaptation performances without the need to manually re-label a new dataset. Although this is a preliminary evaluation, this shows how TTT can significantly reduce the time and effort required for preparing new annotated datasets in medical imaging applications.

This contribution is the base of Chapter 6, and its based on the following publication:

- Colussi, M., Mascetti, S., Dolz, J. and Desrosiers, C. (2024).

ReC-TTT: Contrastive Feature Reconstruction for Test-Time Training. (Accepted for publication @ WACV2025).

My contributions

- Collabotation in concept and methodology design.
- Method implementation.
- Design of the evaluation methods.
- Experiments execution.
- Collaboration in results analysis and interpretation.

1.4.4 System integration

As a more applicative result, we provided a supportive tool for the management of hemophilia, both for patients and practitioners, and combined the contribution described above in a unified system: PRACTICE.

PRACTICE is composed of three main elements:

- GAJA (Guided self-Acquisition of Joint ultrAsound images), an application that provides an automated guiding system to support the patient in the acquisition of joint ultrasound images.
- CADET (Computer-Aided Diagnosis for hEmarThrosis), an application leveraging AI methods to support clinicians in formulating a diagnosis.
- ATOM (Annotation Task Orchestrator Module), a system for the annotation of ultrasound images targeted to clinicians.

In chapter 7 we report on our experience in designing and implementing the system and its components. We also report on the lessons learned in this ongoing project. This contribution is based on the following pubblications:

- Colussi, M., Mascetti, S., Ahmetovic, D., Civitarese, G., Cacciatori, M., Peyvandi, F., ... & Bettini, C. (2023, October). *GAJA-Guided self-Acquisition of Joint ultrAsound images.* In International Workshop on Advances in Simplifying Medical Ultrasound (pp. 132-141). Cham: Springer Nature Switzerland.

- Ahmetovic D., Angileri A., Arcudi S., Bettini C., Civitarese G.,Colussi M., ... & Truma A. (2024, June). *Insights on the development of PRACTICE,a research-oriented healthcare platform.* In 2024 IEEE International Conference on Smart Computing (SMARTCOMP).

My contributions

- Collaboration in concept and methodology design.
- Collaboration in the architecture design.
- Collaboration in the design of the guiding system.
- Collaboration in GAJA implementation.
- Implementation and training of DL models.
- Collaboration in results analysis and interpretation.

1.5 Thesis outline

The remainder of this thesis is organized as follows.

Chapter 2 introduces the medical problem and the relevant state of the art, presenting an in-depth review of current advances in various aspects of data scarcity scenarios for medical imaging. Chapter 3 focuses on the adoption of transfer learning approaches with limited data. We then explore multi-task learning approach in Chapter 4, facing the task of detection and classification of subquadricipital recess distension using US images. In the subsequent Chapter 5, we propose a novel weakly supervised anomaly detection technique specifically designed for US images. Chapter 6 introduces a novel test-time training method that employs contrastive feature reconstruction that is evaluated in various domains, including US images. Finally, Chapter 7 introduces PRACTICE, an intelligent healthcare platform that integrates the methods described above and enables remote monitoring and self-collection of US images by patients, Chapter 8 provides some insight into the experience and impact of the research conducted during the three Ph.D. years and concludes with a summary of contributions, future work, and the general conclusions of this thesis.

2

Background

This chapter describes the problem from the medical point of view and existing CAD solutions for the treatment and care of hemofilia (Sec. 2.1), which will be the base motivation for Chapters 4, 5, 6, 7. Furthermore, we present state-of-the-art techniques addressing the challenges introduced in Chapter 1. Specifically, we explore transfer learning (Sec. 2.3) and multi-task learning (Sec. 2.4) as key approaches to overcoming the issue of data scarcity, anomaly detection, and weak supervision (Sec. 2.5) as a framework for addressing data imbalance, and test-time-training (Sec. 2.6) to deal with domain-shift. Finally, we will present the datasets used for the training and evaluation of the deep learning models described in the various chapters of the thesis (Sec. 2.7). A brief description of all medical terms can be found in Appendix E.

2.1 Hemofilia and joint blood effusion

Hemophilia A and B are rare bleeding disorders caused by a complete or partial deficiency of coagulation factors VIII (FVIII) or IX (FIX). Individuals with severe hemophilia are characterized by low levels of FVIII or FIX, if not treated with adequate prophylaxis, may experience spontaneous musculoskeletal bleeding, which represents approximately 80% of total bleeding events. Recurrent joint bleeding, more frequent in the ankles, elbows, and knees, can lead to hemophilic arthropathy and irreversible joint damage [16]. The improvement in quality of health care and the increased availability of replacement drugs (drugs that replace missing or deficient clotting factor in the patient's blood) and non-replacement drugs (drugs that improve the body's natural ability to clot) have facilitated a change in the objectives of prophylactic treatment. Rather than focusing solely on preventing life-threatening bleeding, the aim of prophylaxis has now shifted towards preserving joint health and improving overall quality of life [17]. Nowadays, in the era of modern hemophilia treatment, including nonreplacement and gene therapy [18], used to preserve long-term joint health, the findings show that efforts should be directed toward early detection of subclinical bleeding, which occurs without visible symptoms, as there is evidence that even a single bleeding episode can lead to cartilage damage and synovitis [19].

Since physical examination is not sufficient to identify joint bleeding by self-report or physician visit [20] ultrasound imaging is emerging as a simple and reliable tool to evaluate joint health in patients with hemophilia. US imaging can identify joint bleeding even in the absence of obvious signs or symptoms, facilitating early intervention and prevention of joint damage. In fact, subclinical joint bleeding is not uncommon due to the availability of both replacement and non-replacement treatments, which mitigate clinical manifestations in patients with hemophilia. Compared to magnetic resonance imaging (MRI), musculoskeletal ultrasound is much more sensitive in detecting bloody (complex) effusion and distinguishing it from synovial effusion and does not involve ionizing radiation. However, its diagnostic sensitivity in differentiating blood clots and synovial hyperplasia can vary depending on factors such as the expertise of the operator and the quality of the equipment used, which could lead to a not completely satisfactory outcome [21, 22, 23, 24, 25, 26]. Recurrent and spontaneous joint bleeding is the most common manifestation of hemophilia, with ankles, knees, and elbows, typically affected and termed *index joints* [16]. Recurrence of intraarticular bleeding fosters the development of synovitis, rendering these joints more susceptible to further bleeds and progressive joint damage.

Ultrasound, performed by a qualified physician, is valuable in assessing distension, blood effusion (indicative of joint bleeding), or synovial hyperplasia, manifested as isoechoic thickening of the capsule [26, 27]. Among the different proposed US scores, early detection of hemophilia arthropathy with ultrasound (HEAD-US) [14] and Joint tissue Activity and Damage Examination (JADE) [28] are applied worldwide. Both aim to assess and score the presence and degree of synovitis and osteochondral damage, although with different definitions of the US image features. None of them defines the characteristics of joint bleeding.

Our research group has recently identified unmet needs in the description of ultrasound features of hemophilic arthropathy and joint bleeding and has proposed a set of definitions as a starting point for their standardization and validation. This process has been proposed and planned as a project in the context of the Factor VIII/IX Standardization Subcommittee of the International Society on Thrombosis and Haemostasis Scientific and Standardization Committee [26].

Ultrasound images. Ultrasound probes uses high-frequency sound waves generated by a transducer that converts electrical energy into sound waves. These waves travel through the body and are reflected by the internal structures. The transducer also detects the returning echoes, analyzing their strength, direction, and arrival timing. These reflected sound waves are then processed to form a grayscale image, with intensity based on the echo's strength. This technique can produce images with high spatial resolution of internal structures of the body, such as tendons, bones, blood, and muscles [12].

The amount of reflected waves depends on the tissue density. Waves that are not reflected propagate to the underlined tissues. Dense structures, such as tendons and bones, produce strong echoes, appearing brighter (echogenic) in the image. In contrast, areas that reflect fewer sound waves, such as fluids, appear darker (anechoic). When ultrasound waves encounter tissues that cannot propagate sound, such as bones, they create a white border and cast an acoustic shadow, making it impossible to detect anything beneath them [29]. An example is shown in Figure 2.1: the patella is clearly distinguishable in light color (see the red box) while the area below it is almost completely black.

Knee subquadricipital longitudinal scan. In the following of this thesis, we focus primarily on one knee joint, more specifically on one of the three scans specified in the HEAD-US protocol for the collection and diagnosis of joint recess distension in patients with hemophilia [14]: the SQR longitudinal scan[1]. This scan is used to assess SQR distension and contains different characterizing elements (see Figure 2.1):

- The femur (blue box) usually appears as a light thick line, approximately horizontal, starting from the left side of the image and extending towards the right, often in the lower half of the image.

- The patella (red box) usually appears as a curved light line, positioned at the right border of the image, often in the top half and not entirely captured.

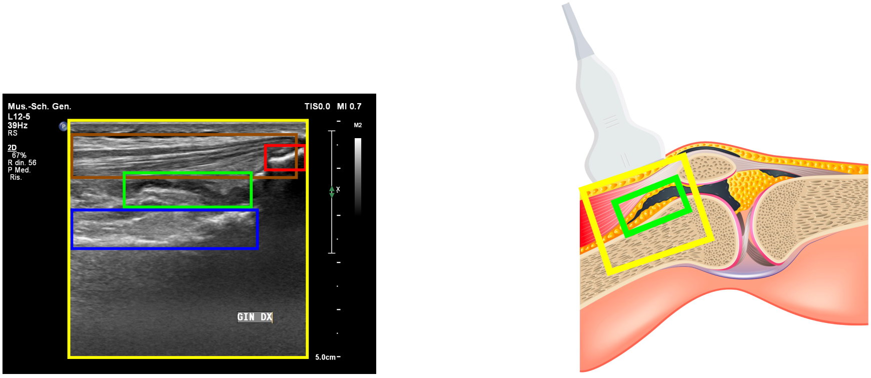
- The quadriceps tendon (brown box) appears as a fascicular structure composed of echogenic parallel lines (*i.e.*, they appear as thin horizontal stripes) that originate from the patella.

[1] A list of all the available scans is reported in Appendix D

The SQR (green box) is positioned between the femur and the patella and often contains at least a small quantity of liquid, hence it is dark. In some cases, the joint recess membrane can be visible in gray. The SQR size and shape vary depending on many factors, including whether it is *Distended* or *Non-distended*, as explained below.

Figure 2.1b shows how the probe must be positioned during the acquisition of the SQR longitudinal scan. In the figure, the yellow box is the area captured by the US image shown in Figure 2.1a, while the green box is the SQR. To correctly acquire this type of image, the knee has to be bent at about 30°. The probe must be positioned right at the beginning of the patella and moved horizontally to identify the correct key features previously described.

Several parameters of the ultrasound probe need to be specified in order to properly acquire SQR longitudinal scans. Some of these parameters need to be personalized for each patient (such as gain, focus and dynamic range), while the value for other parameters can be pre-determined, such as frequency and depth, which in our study were set to $12Mhz$ and $40 - 50mm$, respectively.



(a) Example of SQR longitudinal scan

(b) Probe positioning

Figure 2.1: SQR image acquisition

Elbow olecranic longitudinal scan.

In Chapter 6, we will use the olecranic (OLR) view of the elbow, which is the clearest view of the elbow recess of the HEAD-US protocol. This scan is used to assess OLR distension and contains different characterizing elements similar to the ones of the knee (see Figure 2.2):

- The olecranon fossa (blue box) usually resembles a small, shallow depression or darker area near the distal end of the humerus. It appears towards the lower center of the image.

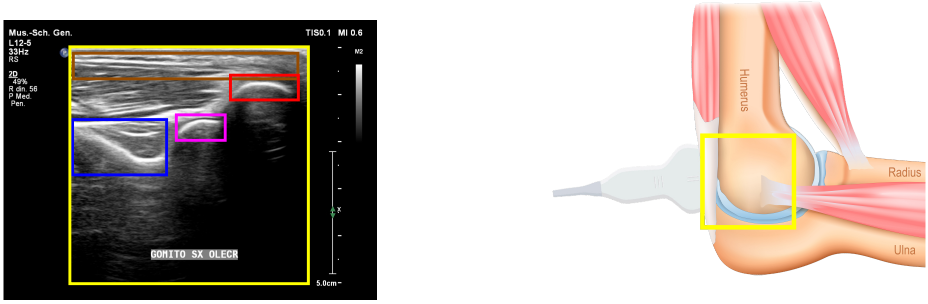
- The humerus (purple box) usually presents as a thick, solid line, roughly vertical or slightly diagonal, starting from the upper left and extending downwards toward the right. It is commonly seen in the upper half of the image, occupying a prominent central position.

- The tricep tendon (brown box) appears as a thin, faintly visible line, running diagonally across the upper part of the image. It originates near the upper right and extends towards the left.

- The ulna (red box) typically appears as a light, thin line, usually oriented horizontally. It begins near the top-right side of the image.

To find the OLR, we first identify the attachment of the tricep tendon and the final part of the humerus bone. Between these two, the recess is positioned inside the olecranic fossa, between two fat pads.

Figure 2.2b shows how the probe must be positioned during the acquisition of the OLR scan. In the figure, the yellow box is the area captured by the US image shown in Figure 2.2a, while the green box is the OLR. To correctly acquire this type of image, the elbow has to be bent at about 90°. The probe must be positioned right at the beginning of the humerus and moved horizontally to identify the correct key features previously described. The same probe parameters described for the knee, need to be adjusted.



(a) Example of OLR longitudinal scan

(b) Probe positioning

Figure 2.2: OLR image acquisition

2.2 Technological Solutions for Hemophilia

Based on the evidence of the importance of promptly identifying and treating joint bleeding to prevent irreversible damage, tools that allow a very early diagnosis of intra-articular bleeding are receiving growing interest. Telemedicine is defined as the distant supply of healthcare services and clinical assistance using information and communication technologies, such as the Internet, wireless systems, and mobile phones. Telemedicine offers several advantages, such as the possibility of treating patients at the point-of-care (POC, *e.g.*, scene of an accident, patient's home) instead of at the hospital, as well as an improved quality of life [30]. However, telemedicine comprises several technologies that need to be addressed, validated, and deployed prior to offering services to patients. Thanks to technological development and the introduction of telemedicine, patients can now use electronic diaries (e-diaries) to record bleeds and treatments (Table 2.1). E-diaries, commonly available as smartphone apps, enable patients to input information about their treatment and bleeding events more quickly and easily than paper diaries. With varying degrees of detail, patients can document bleeds, treatments, and patterns of physical activity.

App	Developer	Platforms	Main Features	Addressability
MicroHealth Hemophilia	Microhealth Inc.	Android, iOS	Logging infusion and bleeding, Set reminders for infusion, Share data via e-mail, manage multiple users, Communicate with hemophilia treatment center via chat, educational resources.	Patients, Caregivers, Doctor
myWAPPS	McMaster University	Android, iOS	Logging infusion and bleeding, Set reminders for infusion, Share data via e-mail.	Patients, Doctors
HemMobile	Pfizer Inc.	Android, iOS	Logging infusion and bleeding, Set reminders for infusion, Share data with physician, Integration with Google Fit and Apple Health.	Patients, Doctors
Florio HAEMO	Florio HAEMO	Android, iOS	Logging infusion, Logging bleeding, Set reminders for infusion, Share data with physicians, Integration with Google Fit and Apple Health.	Patients, Doctors

Table 2.1: Comparison of Hemophilia Management Apps

Teleguidance. The use of portable US imaging systems has been extensively investigated in the literature [31, 32, 33]. Such devices were initially conceived to allow clinicians to make diagnoses at POC. Three approaches have been proposed in the literature. The first approach is to train the patients so that they can independently acquire US images [34, 35, 36]. A different approach is to rely on teleguidance, which means that a medical practitioner remotely supports the patient in real-time during US image acquisition. Teleguidance can be provided by the medical practitioner who observes the US feed (as in [37, 38]), possibly combined with video from other cameras [38]. In [39], the authors suggest that 5G technologies will play a major role in making teleguidance practical in real-world scenarios, and indeed several research groups are exploring it in the general medical domain [37, 38]. One limitation of teleguidance is that it requires the availability of human experts to remotely support the patients or the operator who can perform the ultrasound (US) scans in cases where the patient is unable.

Two recent and closely related study compares these two approaches considering the problem of hemophilic patients using portable probes for self-collection of US images of their joints, with the objective of reducing hospital visits [40, 41]. One of the results of these works is that even if patients follow a dedicated training session (lasting 4-5 hours), the quality of the self-collected images (without any type of real-time assistance) significantly degrades with the passing of weeks, due to forgetting. This suggests that simply training patients is not sufficient for high-quality self-imaging. Interestingly, this work also shows that high-quality images can be self-collected when patients are assisted with teleguidance. However, the problem analysis conducted in our study uncovered that this solution is impractical in our scenario because it is considered to be too time consuming for medical practitioners.

The third approach is based on automated guidance, which means that the patient is guided to correctly position the probe by an AI system. Existing work adopts a camera mounted on the probe to locate and guide the position of the probe [42, 43], and require custom-made hardware devices specifically designed for the problem.

CAD for joint distension with US. Ultrasound [44] is also often used as a data source for Computer-Aided Diagnosis (CAD) systems [45, 46]. In fact, despite its high dependence on the level of operator expertise and possible noise of acquired images [11], US imaging is easily accessible, safe and affordable, and therefore is commonly used in healthcare [46].

In the literature, different solutions have been proposed to automatically detect and classify joint recess distension. For example, a CNN-based method has been proposed to perform segmentation and classification of bicipital peritendinous effusions on the shoulder joint [47]. Specifically, a VGG-16 [48] network is used to extract features and a second CNN is used to classify distension into three classes (*i.e.*, mild, moderate, and severe). The authors evaluated their method on a dataset of 3801 images, including healthy individuals and individuals with BPE with various severity levels, reaching an accuracy of 75%.

Another work considers the knee joints [15] and uses segmentation techniques to classify different types of pathology within US images, including joint recess distension due to synovial thickening.

The authors evaluated the method using 600 US images with 6 different classes (*i.e.*, normal knee joint, non-synovial thickening, synovial thickening, cyst, tumor, rheumatoid arthritis). The results showed an accuracy of $\approx 76\%$.

A closely related work is ARB U-Net which, similarly to our work, extracts Sub-Quadricipital Recess (SQR) of the knee joint from US images [49]. Specifically, ARB U-Net is based on deep segmentation, using an encoder-decoder method that identifies the exact boundaries of the SQR. The results show a segmentation accuracy of 97.1% on a dataset of 450 US images.

There are two main differences between our paper and the three works mentioned above. First, while these studies primarily focus on segmentation-based approaches that require an expert practitioner to precisely annotate the target area, which is both time-consuming and expensive, our work only requires the practitioner to annotate the SQR bounding box, which is much simpler and faster. Additionally, these works do not directly address the issue of recess distension in hemophilic patients, as their focus lies elsewhere.

A recent abstract paper [50] considers US images of patients with hemophilia and addresses the problem of classifying *Distended* and *Non-distended* knee recesses. The authors considered 179 US images collected from pediatric patients, using a CNN to perform binary classification, reaching an accuracy of 82%. Finally, in a study by Ai et al., the authors studied the possibility of predicting the risk of bleeding in a series of children (N = 98) with hemophilia A, using three machine-learning models to evaluate the risk of bleeding during physical activities [51].

A direct quantitative comparison between previous work and our contributions is not possible for two reasons. First, the datasets used for the evaluation of previous work are not public, and hence we cannot evaluate our techniques with the data used in previous work. The second reason is that running existing solutions on our dataset is not possible either, because the first three papers mentioned above require the recess segmentation mask, which we do not have, while the

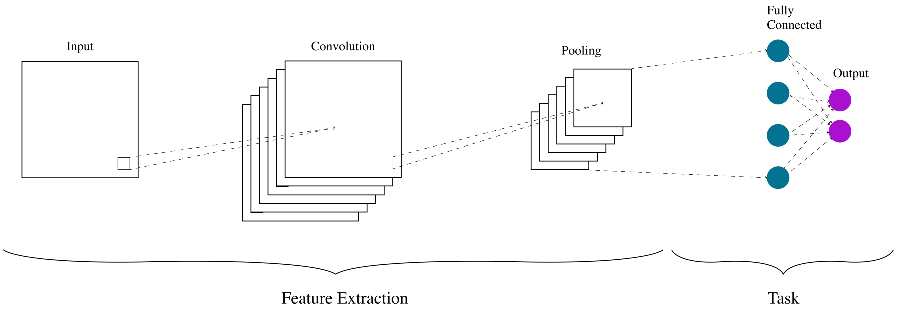


Figure 2.3: Standard convolutional neural network architecture.

last one does not report sufficient details to reproduce the proposed solution.

2.3 State of the art on transfer learning

One commonly used ML approach in US CAD systems is the direct classification of images collected by medical experts [52, 53]. Indeed, different studies adopted deep learning classification approaches to identify various pathologies such as tumors in breast ultrasound [54, 55, 56], liver pathologies [57, 53], thyroid nodules [58, 59], and others [60].

In this problem domain, Convolutional Neural Networks (CNNs) are the most widely used ML architectures, due to their ability to extract discriminative features from image data [61, 48, 60, 62]. CNNs are a specific type of deep learning model designed for the image analysis task. They are inspired by the structures of the human primary visual cortex and can automatically learn and detect spatially invariant features within images, such as edges, textures, and complex patterns. An example of a basic CNN architecture is shown in Figure 2.3

The most common components, that will be referred in the following chapters are[2]:

[2] More specific components, related to only some part of the work will be

- **Convolutional** layers are the main component of CNNs, they are composed of a set of filters, that are trained to capture relations among the input data extracting relevant features. This is achieved by computing the sliding dot product between the input and the filters. The stride parameter controls the amount of slide of the kernel. Padding can also be used to extend the input features in order to control the size of the feature maps. Convolutional layers are staked to extract more general features with the increase of depth.

- **Pooling** layers are used to reduce the dimensionality of the intermediate features of CNNSs. They use a filter that slides along the input features according to their stride. For each iteration, pooling layers applies a function as computing the maximum (maxpooling) or the average (avgpooling) of the input feature map and returns a single value.

- **Fully-connected** are the simplest type of layers, they are composed of a set of neurons, each of which takes the output of the previous layer as input. FC layers are trained to learn the weights to associate with each input.

- **Activation** functions are essential for capturing the non-linear relationships between input data and output, as the preceding layers perform only linear operations. Specifically, they are used to divide whether or not to activate a neuron and, if so, with what intensity. Different activation functions map the input in different ways, for example, the sigmoid activation is monotonically increasing, and it is bounded between one and zero. To overcome problems such as the vanishing gradient, more complex activation functions were introduced, such as the rectified linear (ReLU), which will output zero if the input is negative, while it does not affect positive values.

described in detail in the corresponding sections

The development of deep learning models is often limited by the scarcity of available labeled data for the training of ML models, leading to poor performance [63]. To mitigate this issue, in the literature, transfer learning approaches [64, 58] and generative data augmentation [65, 66] have been proposed.

TL on generic images. Transfer learning is one of the most widely adopted approaches when working with small datasets, primarily due to its ability to leverage pre-trained neural networks that have already learned to extract relevant features from large and diverse datasets. These pre-trained models capture patterns and representations that are broadly applicable across different tasks and domains [67]. After pretraining the network, only a fine-tuning process is typically required to adapt the model to the specific characteristics of the new, less-represented, domain. The fine-tuning step only requires slight adjustment of the weights of the pre-trained model to align them with the new data, rather than training a model from scratch [68]. This is achieved by freezing the feature extraction part of the network, meaning that the weights of these layers are kept constant during training. Figure 2.4 shows the overall framework for transfer learning.

TL has been applied to many different vision tasks, such as image classification [69], object detection [70], segmentation [71], and so on [72].

Transfer learning in medical imaging. As shown in [73] there was a constant growth in the use of TL in the medical imaging field. Its range of applications varies from brain tumor identification [74], lung diseases [75], breast cancer [76], and skin lesion [77], both for classification and segmentation tasks [78, 79], and many in the US imaging domain [77] showing the importance of having large datasets like ImageNet [80] or COCO [81] available for pretraining, even though they represent a different domain.

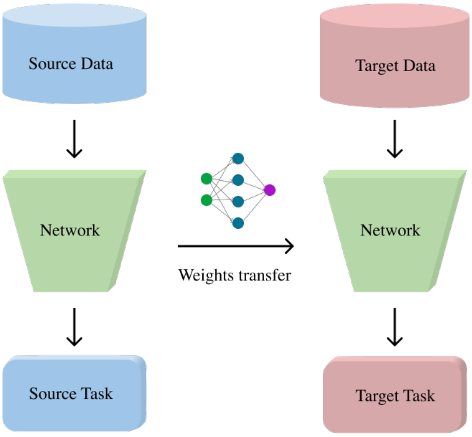


Figure 2.4: Transfer learning approach, the model is firstly trained on a large source domain, and successively updated using the smaller target domain.

2.4 State of the art on multi-task learning

Multi-task learning was introduced to take benefit from the knowledge that a model can learn from one task to enhance the robustness of the other task [82], specifically where the data available for the different tasks are limited. In this framework, the model learns multiple tasks jointly, enabling it to prioritize the tasks it focuses on. In the imaging field, there are two primary types of parameter sharing: hard and soft sharing. In hard parameter sharing, a common feature extractor is shared across different tasks, and the extracted representations are passed to task-specific sub-networks. A combined loss function is used to compute the error for each task, and these errors are summed to optimize all training objectives simultaneously. In contrast, soft parameter sharing involves separate networks for each task, but regularization techniques are applied to keep the parameters of the different networks close, allowing for some degree of shared learning across tasks.

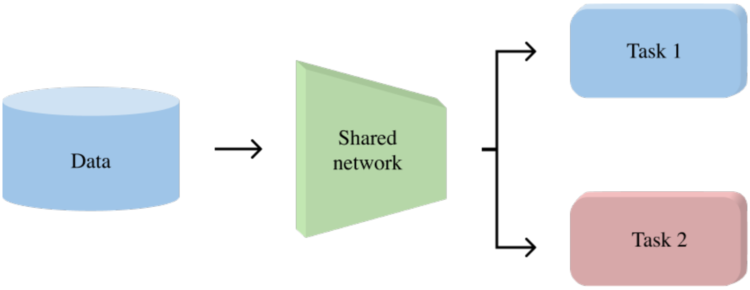


Figure 2.5: Multi-task learning Y-shaped architecture

Multi-task learning in generic images. The more common approach is hard parameter sharing, typically utilizing a Y-shaped architecture, where the shared parameters reside in the lower layers and task-specific layers follow. This method is commonly used in various tasks such as face classification and segmentation [83], segmentation and distance detection [84], iris boundary detection and segmentation [85], fruit cluster and maturity level detection [86], and car color classification, license plate detection, and OCR character recognition [87] and many others [82].

Figure 2.5 shows an example architecture of a Y-shaped network with a shared feature extractor and multiple heads for different tasks.

Multi-task learning in medical imaging. Previous works have explored the multi-task combination of classification and detection for non-US medical images [88, 89, 90, 91, 92], that will be the focus of Chapter 4. A few contributions exploring multi-task learning on US images have also been proposed. Gong et al. propose an approach for multi-task localization of the thyroid gland and the detection of nodules within that region, using a shared backbone network divided into two different decoders for the two tasks [93]. Zhang et al. adopt a multi-task learning algorithm to segment and classify cancer in Breast US images. They propose using *DenseNet121* as the backbone, followed by a decoder branch with layers connected by attention-gated (AG) units to segment the images [94]. The second branch performs a classification task that takes in input the features extracted by the encoder.

2.5 State of the art on unsupervised anomaly detection

Unsupervised anomaly detection (UAD) in computer vision is a technique designed for the specific problem of highly imbalanced datasets, where one of the classes (called *anomalous*) is rare and has few occurrences, such that a standard classification or segmentation algorithm does not correctly understand its characteristics [95]. Its main concept is to learn the distribution of normal data and, therefore, identify when a sample deviates from such distribution, detecting the anomaly. It has been used in many different domains to identify unexpected data, defects [96], fraud [97] or unexpected events [98], leading to prompt identification of potential issues. In the UAD setting, it is assumed that only normal data is contained in the training set, while both normal and anomalous data are present in the testing set [99].

UAD in generic images. The most intuitive approach is using clustering, which leverages the distance within the feature space to differentiate between normal and anomalous data. In this framework, similar images should have close representations in the feature space, while anomalous samples are expected to be more distant from this normal cluster [100, 101].

Other approaches rely on generative adversarial networks (GANs) that can model complex data distributions. Usually, a generator is used to produce synthetic images that are then classified by a discriminator as real or fake. Learning this, the discriminator becomes capable of distinguishing images that do not fall into the learned normality [102]. In addition, conditional GANs have been explored for the anomaly detection task [103].

The main approach for UAD in imaging is the use of autoencoders, a particular type of encoder-decoder architecture designed

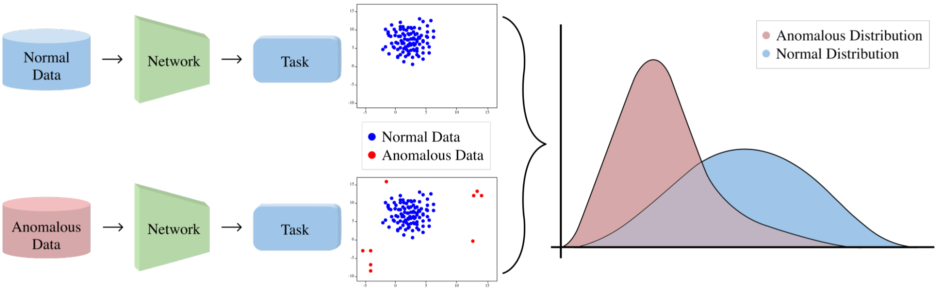


Figure 2.6: Anomaly detection framework, the model is trained on the normal data, then it is used to extract a statistic from the test data, where the anomalies fall outside the distribution.

to first extract a representation of the data (encoding) and subsequently reconstruct the image from that representation (decoding). In such tasks, it is common to define and measure the reconstruction error, usually the difference between the input image and the reconstructed one, which should be higher on anomalous samples, since during the training phase the network is only trained with normal samples [104]. Some methods use inpainting to define more complex reconstruction tasks [104], or use the synthesis of anomalies within normal images [105] or in their feature space [106]. Other approaches use patch-based memory banks [107], and normalizing flows [108].

These techniques make it possible us to obtain, in an unsupervised manner, both an overall anomaly score and a pixel-level anomaly map that can be used for anomaly segmentation. Figure 2.6 shows the standard UAD approach, which consists of training a network and a specific task on normal data, extracting a statistic from the data (blue dots), such as reconstruction error, and finally determining whether a certain statistic falls outside the normal distribution (red dots).

UAD in medical imaging. The rarity of some pathologies and the lack of annotated data highlight the importance of using the anomaly detection framework to learn from normal data, which is easier to annotate and collect. Different works propose specific solutions tackling the problems of medical imaging: normalizing the reconstruction error with uncertainty to detect abnormalities in chest X-ray images [109] or by patch interpolation for brain MRI and abdominal CT images [110], GANs were adopted to detect brain anomalies [111], while more recently the use of denoising diffusion probabilistic models was used for brain anomaly detection [112].

Weak supervision. To address the limited availability of certain labels, a solution known as “weak supervision” emerged, focusing on extracting meaningful information from incomplete or imperfect data. Weak supervision can refer to three different categories: i) incomplete annotations, ii) inaccurate annotations, and iii) inexact annotations [113]. The first refers to the setting in which a small set of clean labels is available and labeling every instance is time-consuming or expensive. This is the usual case of semi-supervised learning approaches [114]. The second refers to the case where annotations are noisy and/or corrupted and therefore cannot be considered gold-standard ground truths. In this scenario, the learning approach exploits labels that might be generated from simpler and less accurate machines [115]. In this thesis, we will focus on the last scenario, the inexact label one. In this scenario, the available data might be labeled, but the labeling is only partial and the given information is not as exact as the desired model output, for example, the coarse region-of-interest of an area to segment [116]. This specific idea has been investigated in multiple works. GrabCut [117] proposed to use the user-provided bounding box to iteratively refine the area to segment inside the rectangles, encouraging neighboring pixels of similar color distribution to have the same label, and more modern techniques such as BoxSup [118] and DeepCut [119] extended this iterative refinement approach using deep learning and CNNs. Some research focuses on the use of weak supervision in anomaly detection, but it is mainly concerned with incomplete annotations [120].

Weakly-supervised anomaly detection in medical imaging. In medical imaging, the significant variability in image characteristics, the presence of noise, and the challenges associated with obtaining large, accurately labeled datasets have led to the adoption of weak supervision in recent approaches to anomaly detection. Some approaches used CAM activation maps of a classification network to refine anomaly maps generated by an unsupervised anomaly detection network on OCT images [121, 122]. Wolled et. al [123] introduce an image-level class annotation to guide a Denoising Diffusion Implicit Model to identify anomalous lung X-ray scans and brain tumor magnetic resonances. Class conditioning is also used to guide a fast non-Markovian diffusion model [124].

2.6 State of the art on domain adaptation

One of the key challenges in deep learning (DL) research is improving the models' ability to generalize effectively to new data. In the most common scenario, models are trained to learn patterns and relations on a source dataset and performance is evaluated on a set of images not seen during training but extracted from the same distribution. Despite the impressive performance achieved by advanced models in various datasets, maintaining the assumption of domain invariance between source and target data proves to be impractical in many real-world scenarios. As a result, the limited robustness of DL models to distribution shifts remains a key obstacle to their use [125, 126].

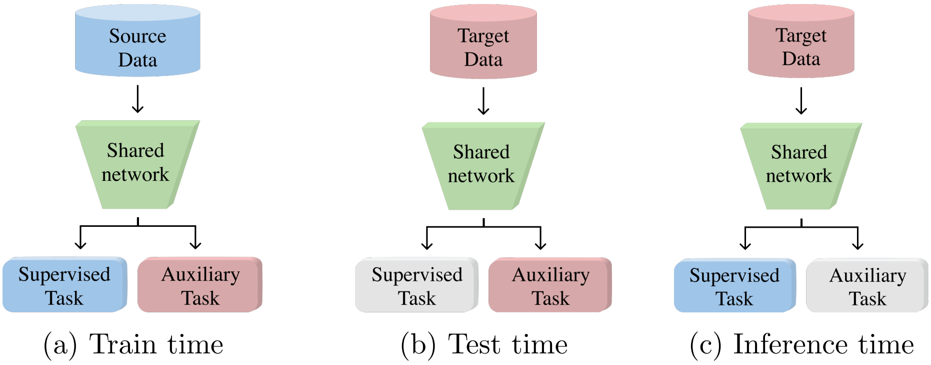
As a solution to this challenge, two broad research branches have emerged: domain generalization (DG) and unsupervised domain adaptation (UDA). DG approaches aim at training more robust models with a native ability to generalize in various domains. The main limitation of these techniques is that they rely on the availability at train time of large amounts of data from different sources, which is often impractical [127, 128]. Moreover, these techniques may also underperform in domains very different from those considered during training.

On the other hand, UDA tries to achieve higher generalizability without anticipating potential distribution shifts but instead by adapting the model accordingly, either with test-time adaptation (TTA) [129, 130] or test-time training (TTT) [131, 132].

Test-Time Adaptation (TTA). In visual recognition tasks, distribution shifts between training and test data can greatly degrade performance [133]. To overcome this issue, recent approaches were proposed to dynamically adapt the models at test time to the new data. Unlike domain generalization [128, 134], where the source model is robustly trained but fixed at test time, TTA allows updating the model for the target domain. TTA techniques do not have access to the source data or training (*i.e.*, only the trained model is provided), and the adaptation occurs only at the test time. In recent years, a variety of TTA methods have been proposed. Among others, in PTBN [135], the adaptation is carried out by updating the BatchNorm layer statistics using the test batch. Instead, TENT [136] tries to minimize the entropy of the predictions for the test set. Finally, TIPI [137] proposes to identify transformations that can approximate the domain shift and trains the model to be invariant to such transformations.

Test-Time Training (TTT). In contrast to TTA, TTT techniques have access to the source data during initial training (but not at test time), and a secondary self-supervised task is trained jointly with the main learning objective. This learning paradigm was first introduced in TTT [131], where the auxiliary task consists of recovering a random rotation of multiples of 90°. At test time, the adaptation is performed by updating only the parameters related to the secondary task. TTT-MAE [138] uses transformers as the backbone of supervised training, with a masked-autoencoder architecture trained as a self-supervised reconstruction task; at test time, the network is trained only to reconstruct the masked images, adapting the shared feature extractor. TTTFlow [139] adopts normalizing flows on top of a pre-trained network to map the features into a simple multivariate Gaussian distribution. At test time, the log-likelihood of this distribution is employed to adapt the model. ClusT3 [132] proposed an unsupervised clustering task that maximizes mutual information between the features and the clustering assignment that should remain constant across different domains.

Figure 2.7 shows an example TTT framework. At train time (2.7a) both tasks are jointly trained. At test time (2.7b) only the unsupervised auxiliary task is used to update the network, while since there is no supervision, the supervised task is not used in this phase. At inference time (2.7c) the main task is used to obtain the final prediction, while the auxiliary task can be discarded.



(a) Train time (b) Test time (c) Inference time

Figure 2.7: TTT framework. (a) During training, the shared network is trained on the source data for both the supervised task and the auxiliary task. (b) At test time, only the auxiliary task is used to update the shared feature extractor. (c) During inference, the model focuses on the supervised task while freezing or disabling the auxiliary task.

Contrastive learning as auxiliary task. Contrastive learning as self-supervised task is gaining remarkable attention in domain adaptation research, due to its ability to learn robust representations. Among the various approaches based on this technique, AdaContrast [140] takes advantage of both momentum contrastive learning and weak-strong consistency regularization for pseudo-label supervision. In *DaC* [141], the test set is divided into source-like and domain-specific, applying two different strategies to the sub-sets using adaptive contrastive learning. TTT++ [142] adopts a contrastive approach on top of a TTT framework, using two augmented versions of the same image as positive pairs and, as negative pairs, augmented versions of other images. This technique also leverages batch-queue decoupling to regularize adaptation with smaller batch sizes. More recently, *NC-TTT* [143] introduces a contrastive approach based on the synthetic generation of noisy feature maps.

In Chapter 6 we propose a TTT methods based on contrastive feature reconstruction, differing from other works, our solution compares the features produced at different network layers, capturing more robust features and allowing for a better alignment in the new domain.

Test time adaptation in medical imaging. The adoption of domain adapation techniques in medical imaging classification usually tackles the change in label distribution, as disease prevalence can vary with location and time. Fang et al. proposed a system for early detection of CT scans of COVID-19 patients, using domain adaptation in the multi-center scenario using a metric-based method finetuning a pre-trained model on the target data using few labels [144]. TTADC [145] proposes to use a set of distribution-calibrated classifiers trained on the source data and, at test-time, aggregate the output of all classifiers using dynamic weights for the different labels.

Most of the research focuses on the segmentation task. Bateson et al. propose a shape-guided entropy minimization loss to adapt a trained model to segment images of a single new patient [146] evaluating the domain shift between the imaging modality change (MRI to CT) and cross-site adaptation. The authors of Adaptive UNet propose a solution for on-the-fly test-time adaptation of a single image by adding an adaptive batch normalization layer to each convolutional block of the network. It is evaluated on different datasets where the shift occurs in sensor acquisition properties, patient age, and resolution [147]. Finally, DLTTA [148] proposes dynamically modulating the amount of weights updated for each test image using memory banks to compute the discrepancy between the source data and the target images. This was evaluated on various classification and segmentation tasks.

2.7 Datasets description

In this section, we provide a comprehensive description of the datasets used throughout this thesis, that are summarized in Table 2.2. This includes both the datasets that were collected, annotated, and preprocessed specifically for the purpose of this research, as well as those that were sourced from established benchmarks for comparative experiments.

Table 2.2: List of datasets used, along with the chapters in which they are referenced and their availability.

Dataset	Chapter	Availability
Knee SQR	4 & 5	Ours, private
Knee and elbow distension	6	Ours, private
CIFAR-10C	6	Publicy available at [149]
CIFAR10.1		Publicy available at [150]
CIFAR-100C		Publicy available at [149]
TinyImagenet-C		Publicy available at [149]
VisDA		Publicy available at [151]
MCSI	3	Ours, available at [152]

2.7.1 Subquadricipital knee recess distension

Despite the fact that there are prior works that analyze US images of the relevant area (SQR scan of the knee) [50, 49, 15], none of these works provides a publicly available dataset. For this reason, we collected a new dataset of 483 SQR longitudinal scan images of 208 adults with hemophilia, aged 44.7 ± 18.6, between January 2021 and May 2022. The dataset was collected thanks to the collaboration with "Centro Emofilia e Trombosi Angelo Bianchi Bonomi" of the Policlinico di Milano, a medical institution specialized in hemophilia. The images were annotated by one expert specifically trained in the diagnosis of SQR distention in hemophilic patients. The study was approved by the institution's ethics committee.

Before acquiring the dataset we first defined a standardized data acquisition protocol that includes: a) examination procedure based on the HEAD-US [14] protocol; b) guidelines on how to use the ultrasound device during the visit, for example defining that the joint side (left or right) should be annotated while acquiring the image itself; c) a procedure for transfering data from the ultrasound device to the hospital server; d) a data pseudo-anonymization procedure.

For each patient, the physician collected several US images from various scans in different joints. For this study, we selected images of the SQR longitudinal scan. Two images of the SQR longitudinal scan are typically collected during each visit, one for each knee (left / right), but for some patients we only have one image while other patients were visited twice (often at a distance of several months), and hence having up to four images each.

Data Acquisition and annotation

Images were acquired using the *Philips Affiniti 50* US device[3] by a single specialized practitioner during routine visits of hemophilic patients. When collecting the images, the probe was positioned as shown in Figure 2.1b and the knee was flexed by 30°. Each image has a resolution of 1024 × 780 and, as shown in Figure 2.1a, it contains acquisition parameters (saved as text in the image) and the actual US scan (*i.e.*, the yellow rectangle in Figure 2.1a), the size of which can vary.

The annotation procedure is organized into three phases. The first

[3] `www.usa.philips.com/healthcare/product/HC795208/affiniti-50-ultrasound-system`

phase is image selection: among all images acquired from the US scanner, those representing the SQR longitudinal scan of the knee are selected. The practitioner discards unsuitable images, such as those of underage patients, of patients with a prosthesis, or images with a wrong knee bending angle. After this phase, a total of 483 images were selected. The second phase is the recess bounding-box annotation. Using an annotation tool [153], the practitioner identifies the position of the SQR and draws the bounding box (a rectangle with edges parallel to the axes).

The third phase is class labeling: the practitioner evaluates whether the recess is *Distended* and enters this information in the annotation tool. Based on this procedure, out of 483 SQR longitudinal scans, 360 were labeled as *Non-distended* and 123 as *Distended*.

Pre-processing

We pre-process the collected images to extract the actual US image (e.g., the yellow box in Figure 2.1a). Indeed, as previously observed [47, 15] using the entire image as returned by the US device can reduce classification accuracy as this part of the image does not contain information needed for the required tasks.

As suggested by Tingelhoff et al. [154], we initially cropped the images manually. However, this process is time-consuming. We therefore developed an algorithm to automatically extract the US scan from the collected image. Figure 2.8 shows the steps of the pre-processing algorithm. In the first step, we measure and binarize the gradient of the image; we then remove connected pixel groups composed of less than 1000 non-zero pixels; afterward, we dilate the image to fill small groups of black pixels, and we perform an opening operation to remove groups of pixels not belonging to the US scan that was merged with it in the previous steps. We cropped the original image with the bounding box of the white area resulting from the previous step. Finally, the images are resized to 256×256 pixels.

All images have been double-checked as part of the annotation process and no cropping error was found, showing that the proposed automatic pre-processing is reliable.

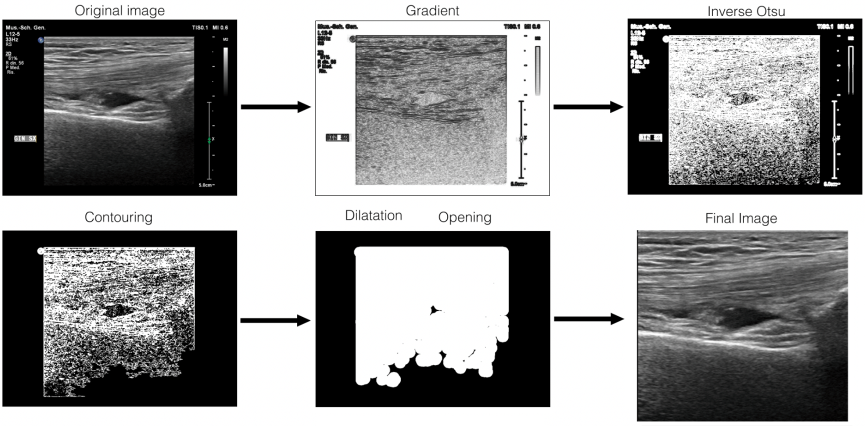


Figure 2.8: Intermediate steps of frame extraction procedure

2.7.2 Knee and elbow recess distenssion

A new recently collected dataset was used to perform a first evaluation of the domain adaptation task on US images. It is composed of two main sets: a source dataset, composed of knee SQR US images, and a target set composed of elbow OLR images. For this dataset, the only available annotation at the time of writing the thesis is the distension annotation, we acquired the annotations of three practitioners and used majority voting to assign the final label to each image.

The knee dataset is composed of a total of 1161, 869 of which are classified as *Non-distended* and 292 as *Distended*.

The elbow dataset is composed of 227 OLR scans, annotated as the knee dataset, resulting in 143 *Non-distended* cases, and 84 *Distended*.

Both datasets were annotated by the same practitioners using the ATOM tool described in Chapter 7.

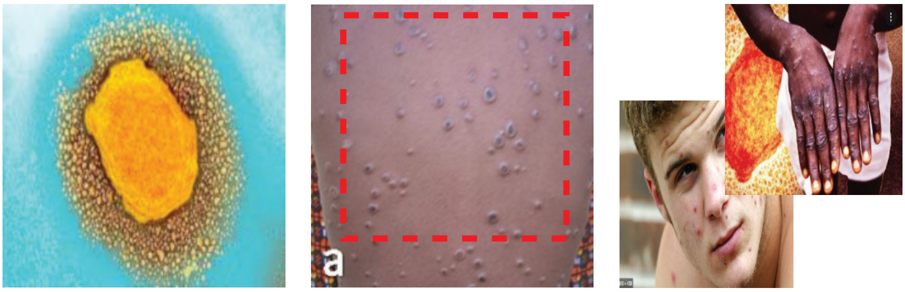
2.7.3 Corruption datasets

Here, we describe the six datasets used to evaluate our domain adaptation algorithm.

CIFAR-10C, CIFAR-100C and TinyImageNet-C[**149**]. These three datasets are composed of 15 different types of corruptions, from various types of noise and blur to weather and digital corruptions. The images present five levels of severity for each perturbation and all the experiments were conducted using only the most severe category (level 5). The datasets consist of $10,000$ test images labeled into 10 classes for CIFAR-10C, 100 classes for CIFAR-100C, and 200 classes for TinyImageNet-C.

CIFAR-10.1[**150**]. We also use the CIFAR-10.1 dataset to evaluate our model's ability to generalize to natural domain shift that takes place when images are re-collected after a certain time. The CIFAR-10.1 dataset is composed of $2,000$ images collected several years after the original CIFAR-10 dataset, with the same 10 classes.

VisDA[**151**]. The Visual Domain Adaptation (VisDA) dataset was designed to pose a new challenge in domain adaptation: from synthetic images to real-world images. This dataset is composed of $152,397$ train images consisting of 2D renderings, $55,388$ validation images extracted from the COCO dataset, and $72,372$ YouTube video frames that compose the test set. All images are labeled into 12 different classes. We evaluated the model's ability to generalize from the training set to the validation set ($train \rightarrow val$) and from the training set to the test set ($train \rightarrow test$).



(a) No skin sample (b) Cropped (c) Whole body parts

Figure 2.9: Examples of the criteria applied during the dataset creation.

2.7.4 The Mpox Close Skin Images dataset

The creation of a skin image dataset was essential to begin evaluating deep learning techniques in the context of limited medical images. Skin conditions often provide a more accessible and varied set of images compared to other medical fields, allowing a preliminary study for training models with scarce data. We therefore introduce Mpox Close Skin Images (*MCSI*), a dataset that has been created according to three design principles. First, the dataset only includes close skin images with or without skin lesions, as these are representative of the pictures that users can collect in the use case considered. Second, *MCSI* contains images of skin lesions caused by diseases that, according to the WHO, should be considered in the clinical differential diagnosis of mpox [155]. In particular, we consider one class for chickenpox rash and one for acne, which is a common skin condition caused by bacterial skin infections. Third, the number of samples is balanced among the different classes to avoid bias.

Specifically, *MCSI* includes: (1) images of **Mpox** cases collected by Ali et al. [156] by web scraping news portals, publicly available case reports, and websites; (2) pictures of **Chickenpox** lesions available on the Hardin Library for the Health Sciences of the University of

Iowa[4], (3) samples of **Acne** at different severity levels, collected by Wu et al. [157] and freely available on Github[5], and (4) samples of skin without evident lesions, named as **Healthy**, available in the dataset collected by Muñoz-Saavedra et al. [158].

In order to create the MCSI dataset we followed a two-step procedure: first, we excluded images where no skin is visible (as in Figure 2.9a). Then, for the remaining images, we selected the larger square area that contains the skin and no background (see example in Figure 2.9b). The area is discarded if its sides are less than 224 pixels long. This is due to the fact that some original images contain whole body parts (as in the examples shown in Figure 2.9c) and hence the selected area can result in low resolution.

Currently, MCSI dataset labels are derived from those available online, and no verification has been conducted by expert medical practitioners. However, we intend to verify the validity of the annotations in MCSI with the collaboration of medical experts as part of our future work.

The resulting dataset comprises a total of 100 images for each of the 4 designated categories. Figure 2.10 provides a representative selection of images from our dataset, showcasing examples from each category. The dataset has been made publicly available [152].

[4] `http://hardinmd.lib.uiowa.edu/chickenpox.html`

[5] `https://github.com/xpwu95/LDL`

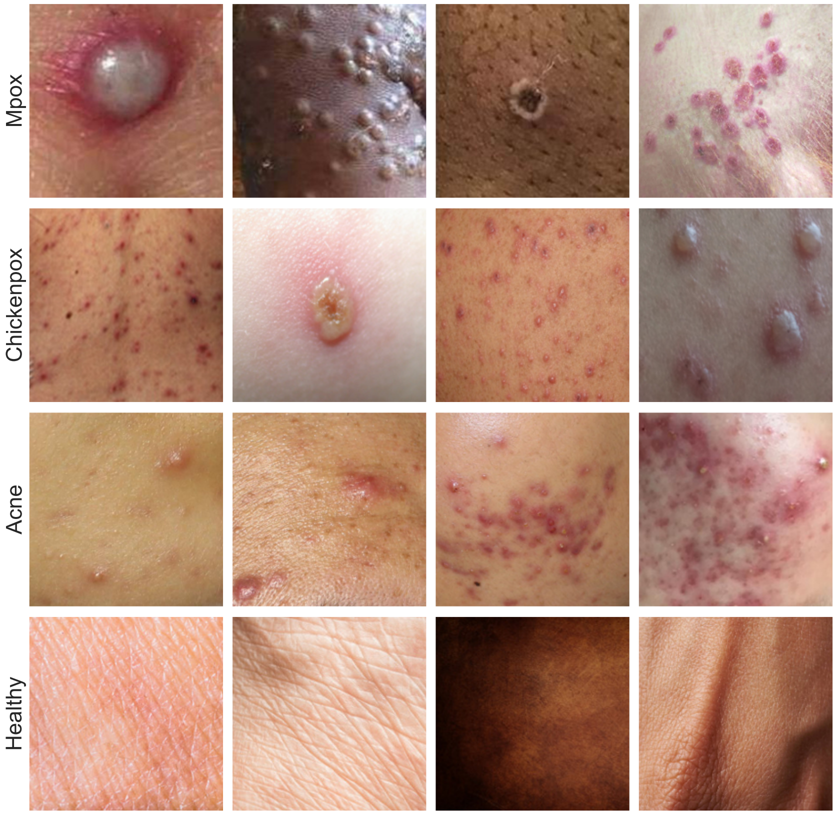


Figure 2.10: Sample images from the collected dataset for each of the 4 considered classes: *Mpox*, *Chickenpox*, *Acne*, and *Healthy*.

3

Adoption of Transfer Learning Approaches to Detect Mpox using Smartphone images

While the whole world is still dealing with the coronavirus disease (COVID-19) and its mutations [159], the recent outbreaks of mpox[1] virus (formerly known as Monkeypox) in different western countries have raised serious concern among public health authorities [160]. The mpox is a zoonotic disease caused by an orthopoxvirus, and it is closely related with variola (i.e., the smallpox virus), cowpox, and vaccinia viruses [161]. Although it was first isolated in 1958 from laboratory monkeys, its original hosts also included squirrels, rats,

[1] In the rest of the chapter we will use Mpox with capital letter when referring to the detection class, while mpox when referring to the virus.

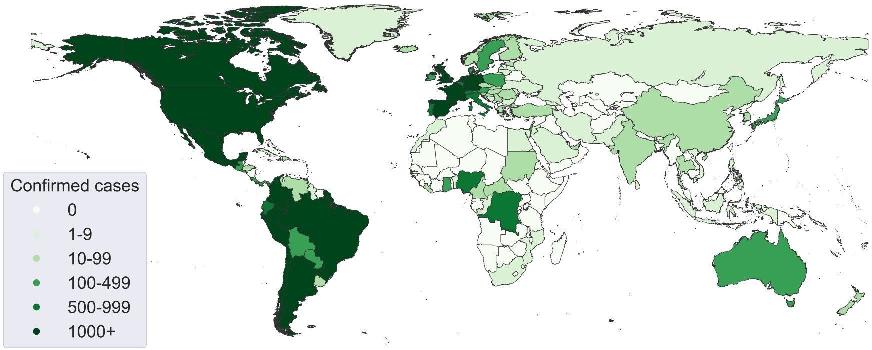


Figure 3.1: Geographical distribution of the recent mpox outbreak [7].

and dormice [162].

Since the first human case reported in 1970 in the Democratic Republic of Congo, the spread of mpox has been always limited to Central and West Africa, infecting new hosts through close body contact, respiratory droplets, or animal bites, becoming an endemic disease in those regions. The incubation period ranges from 5 to 21 days, and the actual disease is characterized by generic symptoms such as fever, intense headache and muscle pain, while the most characteristic sign of mpox is related to the appearance of skin rashes and eruptions that usually begin within 1–3 days of the appearance of fever and tend to be more concentrated on the face and extremities rather than on the trunk [163].

Since the middle of 2022, a continuously increasing number of cases and sustained chains of transmissions have been reported in regions without direct or immediate epidemiological links to endemic areas, including countries in Europe, North America, and Australia. On 19 September 2023, the World Health Organisation (WHO) reported a total of 90,465 laboratory confirmed cases and 663 probable cases across 115 countries [7], as shown in Figure 3.1. Even though mpox is usually not fatal, according to the Centers for Disease Control and Prevention (CDC), people with severely weakened immune systems, children under 1 year old, subjects with a history of eczema, and pregnant or breastfeeding women may more likely get seriously ill or even die [164].

Such rapid and widespread dissemination of the virus has raised several worries in the medical community, highlighting the need for proactive countermeasures in order to prevent another global pandemic [163]. In this regard, recent studies have emphasized how *mobile-health systems* (*m-health*), along with Artificial Intelligence (AI), can represent a game changer in containing the spread of a virus [165, 166]. In fact, using the plethora of sensors embedded in modern mobile devices and their increasingly advanced computational capabilities, smartphones and wearables can be used as low-cost, pervasive, and non-invasive tools to support the early diagnosis of new cases. For example, Rong et al. developed a smartphone-based fluorescent lateral flow immunoassay for the detection of Zika virus [167], Brangel et al. proposed the use of a mobile application to read immunochromatographic strips to detect antibodies against Ebola [168], while more recent works used Deep Learning (DL) models to detect COVID-19 digital biomarkers in respiratory sounds collected by smartphone microphones [169, 170]

In this chapter, we propose a DL-based m-health solution to detect mpox from skin lesion images captured by personal smartphones. The considered use case is the following: the user takes a close picture of a skin region that the application uses to automatically detect mpox. Technically, we use Transfer Learning [171] to adapt state-of-the-art Convolutional Neural Networks (*CNNs*) models [172] to automatically identify visual features of mpox skin rashes, distinguishing the typical symptoms of the virus from skin lesions produced by other pathologies that can be easily confused also by expert eyes, including Chickenpox and Acne, at different severity levels.

Compared with previous works, this work addresses three issues. First, the elaboration of available skin lesion images to make them homogeneous with respect to skin section focus and measure, to generate a new homogeneous dataset. In fact, existing datasets include highly heterogeneous images (*e.g.*, images of a group of people or of entire parts of body) that are unsuitable for the considered problem.

Second, the design of a mpox detection system able to run autonomously on personal mobile devices at least to provide a preliminary warning to common users, and that relies on cloud components only for model training and interaction support with a medical expert. To this end, we *optimize* the final DL model to reduce by 4× the memory footprint of our system, without negatively affecting its classification performance.

Third, the integration of *eXplainable AI* (*XAI*) methods [173] to validate the system performance in recognizing the disease from skin lesion pictures and further define a clinical validation process involving medical experts. According to the literature, XAI techniques greatly improve the general understanding of deep neural networks [174], increasing the trust in the overall system by both medical personnel and final users, thus fostering widespread adoption of such digital solutions. In fact, the target of our proposal is twofold: on the one hand, medical experts can take advantage of such a tool to speed up the diagnosis of new cases, while, on the other hand, final users can autonomously perform a preliminary screening of suspicious skin lesions that must be further investigated by their personal physicians or dermatologists.

3.1 Mpox detection system for mobile devices

Figure 3.2 shows the high-level architecture of the proposed framework to detect mpox from skin lesion images collected from mobile devices. The whole process can be summarized in two main stages. In the first stage, we rely on the Transfer Learning approach to adapt a set of pre-trained CNNs to our application scenario, using MCSI to fine-tune their parameters. The rationale for using existing CNNs is that they have been proven to be effective in addressing classification problems in the medical imaging domain [175]. However, one limita-

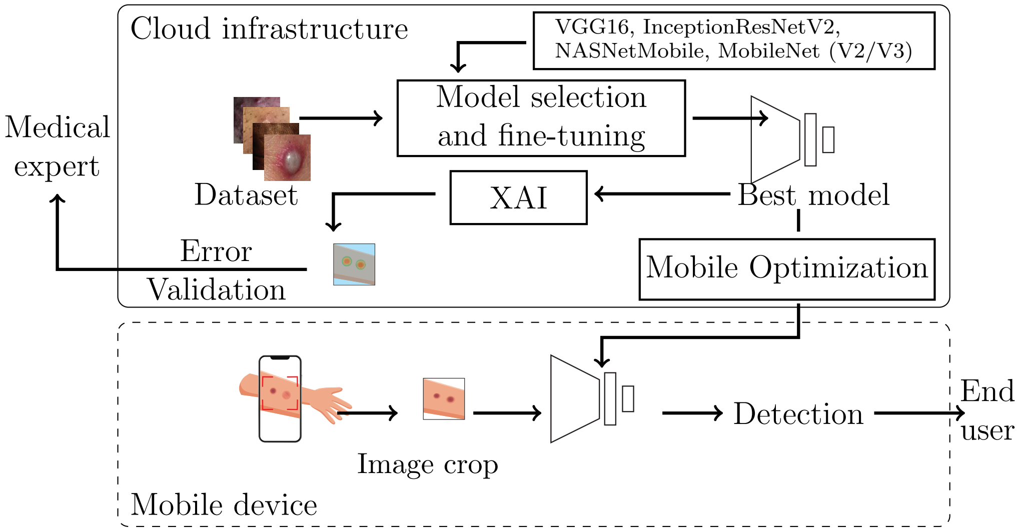


Figure 3.2: Scheme of the mpox diagnosis infrastructure considered in this work.

tion of the CNNs is that they need to be trained on a large amount of data (e.g., Imagenet[80]) and this is extremely expensive in terms of computational time and resources. We address this limitation by using existing CNNs for which pre-trained weights are available.After the experimental comparison of the models' performance, we identify the best model for our mpox detection system, which is then optimized for mobile devices. Since the fine-tuning process includes complex and time-consuming operations, it is executed on a remote server.

The second stage involves the use of the optimized best-performing model to identify new mpox cases, performing the whole data processing on user devices: a new picture is firstly acquired from the device camera and then cropped in order to contain the target skin lesion. The resulting image is then used as input to the deep learning model that generates the classification. Moreover, a XAI module is used to both explain and, to some extent, validate the model's prediction, highlighting the most important sections of the input image that led to the model output.

In the following, we describe in detail the main building blocks of the proposed solution.

3.1.1 Model selection and fine-tuning

The framework relies on transfer learning to adapt a set of pre-trained CNNs to our application scenario, thus reducing the dependence on a large number of training data to build up the target learners [171].

We consider the following 5 CNNs that represent the state-of-the-art on image classification:

- **VGG-16** [48], composed by 5 consecutive blocks of convolutional layers for features extraction, followed by 3 fully-connected layers for classification. Convolutional layers use 3 × 3 kernels with a stride of 1 and padding of 1 to ensure that each activation map retains the same spatial dimensions as the previous layer. A Rectified Linear Unit (ReLU) activation is performed right after each convolution, and a max pooling operation is used at the end of each block to reduce the spatial dimension. Max pooling layers use 2 × 2 kernels with a stride of 2 and no padding to ensure that each spatial dimension of the activation map from the previous layer is halved. Finally, two fully-connected layers with 4096 ReLU activated units are used before a final 1000 fully-connected softmax layer.

- **Inception-Resnet-V2** [176] represents a combination of two popular architectures: GoogleNet [177] and ResNet [178]. While the former is based on the concept of "Network in Network" [179], where a large number of convolutional kernels constitute a very deep architecture to increase the network's generalization, the latter introduced the idea of directly bypassing the input information to the output, thus changing the direct learning target value into learning the residual value between the input and the output. Inception-Resnet-v2 combines the two concepts, using residual connections instead of filter concatenation, to both accelerate the training and improve the performance.

- **NASNetMobile** [180], a simplified version of Neural Architecture Search Network (NASNet) proposed by GoogleBrain, which is a scalable CNN architecture consisting of basic building blocks, called *cells*, that are optimized using reinforcement learning. A cell consists of only a few operations, including both convolutions and pooling, which are repeated multiple times according to the required capacity of the network. The mobile version consists of 12 cells, with a total of 5.3 million parameters.

- **MobileNetV3** [181], a CNN-based architecture especially tuned to best performing on smartphone CPUs through a hardware-aware Network Architecture Search (NAS), combining a series of building-blocks developed by previous models: the depth-wise separable convolutions as an efficient replacement for traditional convolution layers from MobileNetV1 [182], the linear bottleneck and inverted residual structure introduced by MobileNetV2 [183], and the lightweight attention modules used in MnasNet [184]. The model comes in two flavors - which both are tested in this work - that are **MobileNetV3-Large** and **MobileNetV3-Small**, which are targeted for high and low resource use cases, respectively.

For all the aforementioned architectures, we take into account their instances pre-trained with ImageNet [80], a large-scale dataset of 3.2 million images and 1000 different labels, which is commonly used to train CNNs in the image classification domain [185]. Note that ImageNet does not contain labels related to the specific problem domain considered in this chapter. To mitigate this domain shift, we employ Transfer Learning replacing the last fully-connected layers of the network with a novel set of classification layers fine-tuned with MCSI dataset.

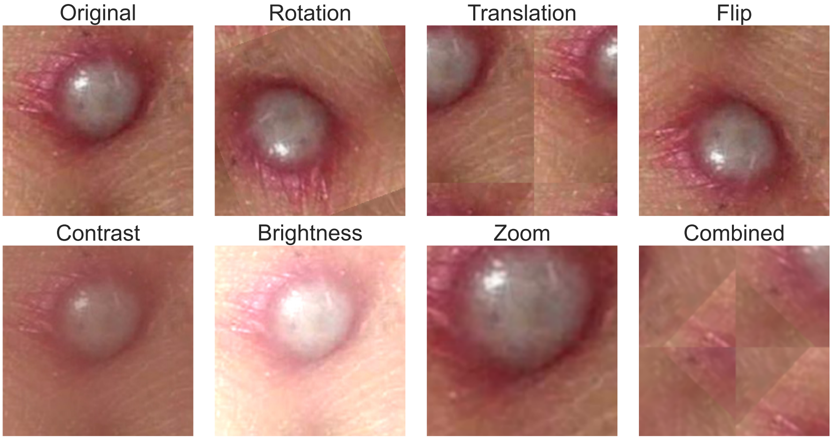


Figure 3.3: Example of data augmentations used in our experiments.

We then validate the considered models through the use of a 10-fold cross-validation procedure and *Hyperband*, a broadly used hyperparameter selection algorithm for deep neural networks, which is able to speed up the random search over the parameter spaces through adaptive resource allocation and early-stopping [186]. In other words, Hyperband uses a combination of small random searches aimed at partitioning the original search space into smaller sub-spaces. Once a search iteration is completed, the most promising sub-spaces (i.e., those that allowed the network to obtain the best results) are further explored until a performance plateau is reached or the iterations budget (i.e., the maximum number of iterations) has been exhausted. In this process, we exclusively fine-tune the final classification layers, which drastically decreases the number of parameters to be trained and, consequently, the amount of data required for the training. Furthermore, to mitigate the risk of model overfitting during the training phase, we employ standard techniques, including *Early Stopping* and *Dropout*.

Furthermore, during the evaluation process, we investigate the feasibility of using data augmentation in our application scenario to possibly improve the performance of the fine-tuned models. Specifically, we employ the 6 standard image augmentation techniques [187] shown in Figure 3.3: (i) *Rotation*, which changes the image angle, simulating different orientations; (ii) *Translation*, simulating different positions of the skin rash inside a specific picture; (iii) *Flip*, which mirrors the image, thus simulating different type of pimples; (iv-v) *Contrast* and *Brightness*, simulating different settings in the amount and intensity of light; and, finally, (vi) *Zoom*, scaling the image to simulate variations in the distance between the skin lesion and the smartphone camera.

Data augmentation is not applied to the test and validation sets to avoid introducing bias in the models' evaluation. We include the parameters that affect the augmentation factors (e.g., rotation angle or zoom level) into the tuning phase to identify the set of values that lead to the best classification performance for our application scenario.

3.1.2 CNN optimization for mobile devices

Our main goal is the definition of a mpox detection system that can be entirely executed on mobile devices. However, neural networks are both computationally and memory intensive. While modern smartphones are equipped with increasingly powerful hardware (e.g., multicore CPUs and, in some cases, dedicated GPUs) that allows performing the inference phase in just a few milliseconds, neural models' size still represents a challenge, making it difficult to deploy them on embedded systems with limited memory resources.

To cope with this issue, several techniques have been recently proposed to reduce the memory footprint of deep learning models, including *pruning*, where redundant connections among hidden units are removed, or *weight clustering*, which consists in replacing similar weights in a layer with a representative value found by clustering algorithms [188, 189]. *Quantization* is another practical and broadly used technique to optimize deep learning models by simply lowering the operations' precision from 32-bit floats to 16-bit floats or even 8-

bit integers. Despite its simplicity, it is generally effective in reducing the overall model's size by 4× at least, with little or no degradation in terms of accuracy [190]. Furthermore, while other approaches must be used during the training phase, quantization can be applied to the final fine-tuned model yield by transfer learning.

3.1.3 Explaining the model's predictions

Deep learning models including CNNs are weak in explaining their inference process and final predictions, thus being typically considered as a black-box. This characteristic is not suitable for many real-world applications, and especially for the health sector, in which explainability and transparency are essential not just for researchers and developers to validate their models, but also for the users who can be directly affected by AI decisions.

For this reason, increasing attention has recently been paid to eXplainable AI (XAI) techniques with the aim of making AI models more transparent, understandable, and interpretable, so as to increase trust in their predictions. Different XAI approaches have been recently proposed for deep learning models, based on the characteristics of specific architectures [173]. According to Ibrahim et al. [191], XAI techniques for CNNs can be categorized as *decision models* and *architecture models*. While the former solutions aim at identifying the parts of an image that mostly contributed to the network decision, the latter explore the network internals, analyzing the mechanism of both hidden layers and neurons.

Given its simplicity in both implementation and interpretability, for our mpox detection system, we decided to use Grad-CAM [192] as XAI approach, one of the most popular decision models used in medical imaging [193, 194]. Grad-CAM is defined as an importance attribution feature algorithm that generates a visual explanation for class-discriminative prediction. Specifically, it captures the features that positively influence the prediction of a given class, by computing its gradient and then propagating it back to the last convolutional layer to finally generate a heatmap that visually represents the most relevant part of the input image that has led the model to that prediction. As a preliminary stage, this approach represents a useful tool to validate the ability of the considered fine-tuned deep models in correctly detecting mpox. Then, after a thorough clinical validation performed by experts with a larger amount of data, such a XAI technique might be also implemented on the mobile device of the final user to support the pre-screening of suspicious skin lesions.

3.2 Experimental evaluation

In this section, we present the experimental evaluation performed to identify the best DL model. We first describe in detail the evaluation protocol and metrics adopted to measure the classification performances of the fine-tuned CNN models. Finally, we discuss the obtained results.

3.2.1 Evaluation protocol and metrics

The evaluation protocol is based on the following: we decided to rely on *10-fold stratified cross-validation* to avoid biasing the results based on specific train/validation/test splits of the dataset. The procedure can be summarized as follows. Firstly, we partition the dataset into 10 folds, ensuring that all the considered classes of images are equally represented in each fold. For each of the 10 cross-validation iterations, one fold is selected as the *test set*, while the remaining 9 represent the *development set* that is further divided into stratified non-overlapping *train* (75%) and *validation* (25%). We apply data augmentation at run-time, only on the training sets. Then, a hyperparameters tuning process (Section 3.2.2) is used by training models on the train set and testing them on the validation set. The model yielding the best performance is then tested on the test set, providing the performance for that iteration.

We measure the average performance of the fine-tuned models obtained during the 10-fold cross-validation by using the different base models as backbone for features extraction, and a set of fully-connected layers are trained from scratch for classification. We consider the following standard classification metrics: *Accuracy*, which is the percentage of correct predictions; *Sensitivity*, which represents the true positive rate; *Specificity*, that indicates the true negative rate; and *F-1 Score*, which is the harmonic mean of Precision and Sensitivity.

We perform the whole process for two different classification settings: binary and multiclass. In the former, we evaluate the models' ability to identify mpox cases without distinguishing the other classes, which are merged into a single "other" class. Since in this setting the training data are unbalanced, we replace the standard F-1 Score with its micro average in order to avoid biasing the results towards the majority class (i.e., "other"). By contrast, in the latter setting, the models learn to distinguish all the four classes available in MCSI.

Furthermore, we conduct a statistical analysis to determine the level of significance in the obtained classification results in terms of accuracy, thereby identifying the most effective model(s) for our specific application scenario. Initially, we examine the outcomes of the two classification tasks without employing data augmentation. We conduct this analysis by using *Repeated Measures Analysis of Variance (ANOVA-RM)*, a statistical method used to assess significant differences among the means of three or more dependent groups. We chose this method because our models were evaluated on the same data folds, making the results dependent on each other. Moreover, even though ANOVA is generally robust to slight deviations from normality assumptions (especially with small sample sizes), we use the *Shapiro-Wilk* test to assess the distribution characteristics of the results. This evaluation aimed to confirm that the models' results can be approximated by a normal distribution. Since ANOVA-RM only indicates the presence or absence of a significant difference, without specifying the specific groups that differ from each other, we subsequently employ the *Tukey's Honest Significant Difference (HSD)* test, which allows us to determine the significance of performance differences between each pair of models, providing a more detailed understanding of the disparities.

Next, we perform a statistical assessment to evaluate the impact of data augmentation on each model, by employing the following procedure. The initial step involves using the Shapiro-Wilk test to determine whether the performance of the model, both with and without augmentation, follows a normal distribution. If both distributions pass the test (i.e., $p > 0.05$), we proceed to assess their homoscedasticity using *Bartlett's* test, which determines if the distributions have equal variances. However, if either distribution failed the Shapiro-Wilk test, indicating non-normality, we utilize the non-parametric *Wilcoxon's rank-sum* test as an alternative to the two-sample t-test. Finally, if the distributions exhibited homoscedasticity, we employ the standard *Independent t-test* to evaluate their statistical significance; otherwise, we use the *Corrected Independent t-test* (also known as *Welch's test*) instead.

3.2.2 Hyperparameters tuning

Actual performances of deep neural networks depend on several hyperparameters that must be tuned in order to find the best configuration for every application scenarios. We adopted Hyperband for fine-tuning the model and data augmentation parameters. Considering the model's parameters, we tune the *learning rate* (`LR` in the range $[1e-6, 0.001]$) and the *number of classification layers* (`N_layers` among values $\{1, 2, 3\}$). Then, for each classification layer, we tune the *number of hidden neurons* (`Dense` among the values $\{256, 512, 1024, 2048, 4096\}$) and the *dropout* rate (`Dropout` in the range $[0, 0.5]$).

Regarding the data augmentation, we explore two different types of parameters' spaces: continuous and discrete. The former is defined within $[0, 0.5]$ and governs the application of `Rotation`, `Zoom`, `Contrast`, `Brightness`, `Translation` (both horizontally, `Tr-width`, and vertically, `Tr-height`), indicating the percentage in which each operation is applied on the original image. For example, the value 0.2 for `Rotation`, represents a random rotation of the image between

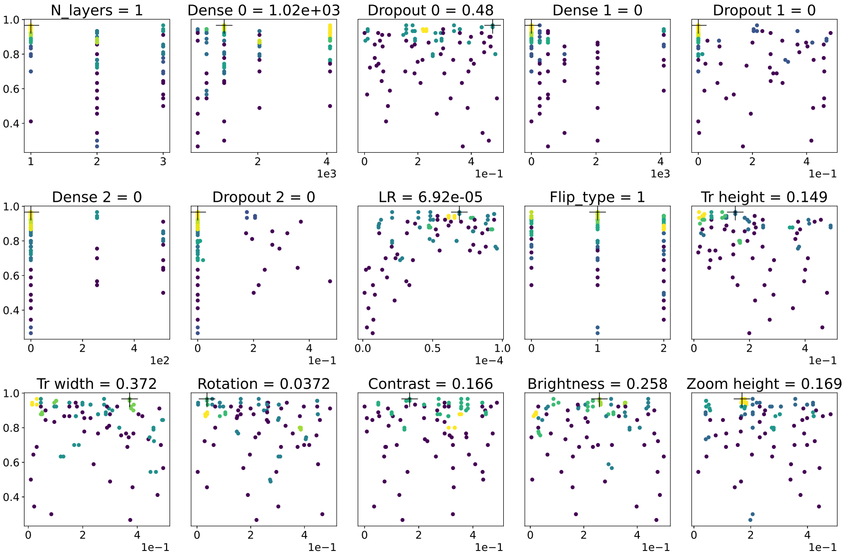


Figure 3.4: Explored parameters for MobileNetV3Large with augmentation (on fold 0)

[−20%, +20%]). The latter controls the application of `Flip type`, which may be applied in three different modalities: *Vertical* (0), *Horizontal* (1), and the combination of the two (2).

Figure 3.4 shows an example of the parameters space explored by Hyperband during the fine-tuning of MobileNetV3Large with data augmentation. The X-axis indicates the exploration space for a given parameter and can include a finite set of values (*e.g.*, the `N_layers`) or can be continuous in a given interval (*e.g.*, `Dropout`). Instead, Y-axis indicates the accuracy levels. In order to ease the visualization, the density of points is shown with colors (with the *viridis* color map): a single point is shown in purple while multiple overlapping points are shown in yellow. Finally, the cross symbol (+) highlights the combination of parameters that produced the best results, which is also reported on the sub-plot titles. Note that the parameters `Dense` and `Dropout` refer to the corresponding classification layer.

Table 3.1: Binary classification performance of the considered base models, with and without data augmentation in the training phase. The performance is reported as mean and standard deviation over the 10-folds of the cross-validation.

Base model	Augmentation	Accuracy	Sensitivity	Specificity	F-1 Score
VGG16	✗	.898 (±.059)	.833 (±.106)	.710 (±.223)	.897 (±.057)
	✓	.890 (±.028)	.835 (±.027)	.730 (±.067)	.890 (±.028)
InceptionResNetV2	✗	.732 (±.051)	.568 (±.063)	.240 (±.196)	.734 (±.052)
	✓	.728 (±.068)	.544 (±.109)	.180 (±.244)	.728 (±.068)
NASNetMobile	✗	.811 (±.038)	.726 (±.061)	.550 (±.151)	.812 (±.037)
	✓	.835 (±.044)	.727 (±.080)	.510 (±.173)	.835 (±.044)
MobileNetV3Small	✗	**.930**(±**.041**)	.877 (±.067)	**.780**(±**.123**)	**.929**(±**.040**)
	✓	.921 (±.043)	.872 (±.062)	**.780**(±**.114**)	.919 (±.040)
MobileNetV3Large	✗	.930 (±.042)	.861 (±.086)	.730 (±.177)	.928 (±.040)
	✓	**.930**(±**.039**)	**.878**(±**.071**)	**.780**(±**.140**)	.928 (±.037)

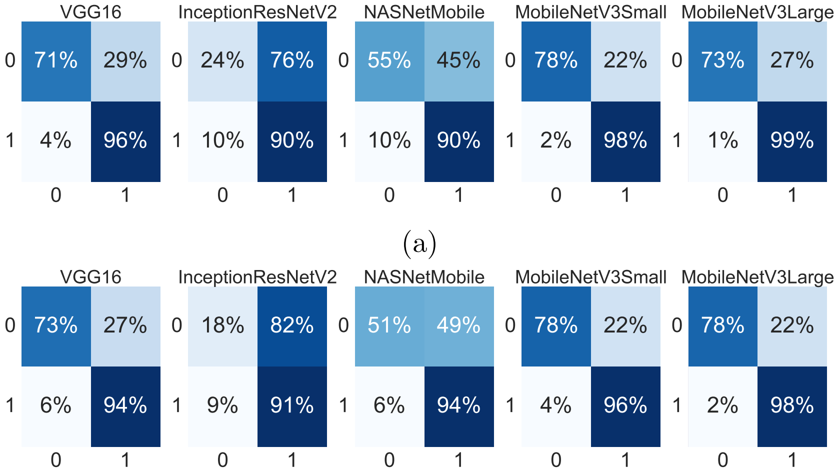
So, for example, `Dense 1` represents the number of hidden neurons in classification layer 1. Hence, if a classification layer does not exist (as in the case of layer 2 when `N_layers` is 2) the corresponding `Dense` and `Dropout` parameters have a value of zero.

3.2.3 Mpox detection performances

In this section, we present in detail the results obtained by fine-tuning the considered CNN architectures in both binary and multiclass classification settings, with and without data augmentation. We also present an analysis of their ability to correctly represent image data samples in the latent features space, thus providing additional support to the standard evaluation metrics.

Binary classification task

Table 3.1 summarizes the binary classification results of the fine-tuned models, both with and without data augmentation; the results are expressed in terms of mean and standard deviations of the considered evaluation metrics, calculated over the 10-folds of the cross-validation.



(a)

(b)

Figure 3.5: Confusion matrices related to the binary classification task with original training data (a) and by employing data augmentation (b). Label 0 refers to `Mpox` samples, while label 1 indicates the generic class `Others`.

Most of the considered base models are able to reach an accuracy level above 80%. InceptionResNetV2 performs worst, thus clearly indicating that such an architecture is not able to detect mpox skin rashes from lesions produced by other pathologies. This is even clearer by observing the confusion matrix in Figure 3.5, noting that the model incorrectly classifies 76% of the overall Mpox samples with the original training data and 82% with data augmentation.

NASNetMobile obtains better results than InceptionResNetV2, but its specificity score is still too low, and its misclassification rate is particularly high to be considered a valid candidate for our system. On the other hand, VGG16 performs better than the previous models. In this case, we can also note a small improvement introduced by using data augmentation, reducing the percentage of incorrectly classified mpox samples from 29% to 27%.

The two variants of MobileNetV3 obtain the best results, reaching in both cases an average accuracy level of 0.93 and with comparable results for all the considered metrics. MobileNetV3Small is able to reach the maximum value also in terms of F-1 score, overcoming by approximately 10% the performance of the larger model. In terms of misclassification rate without data augmentation, MobileNetV3Small improves MobileNetV3Large by 5%, while the larger model performs slightly better in classifying data samples labeled `Others`. On the other hand, in this case, data augmentation seems to introduce more confusion in the model predictions. In fact, while it allows MobileNetV3Large to improve its `Mpox` detection rate, at the same time, it increases the misclassification of `Others` samples for both models, reaching an error rate of 4% and 2% for MobileNetV3Small and MobileNetV3Large, respectively.

Despite MobileNetV3 achieving the highest classification score, the statistical analysis does not reveal significant differences in accuracy compared to VGG16, with a probability of $p = 0.609$. On the contrary, the analysis confirms that InceptionResNetV2 is the least performing model, exhibiting lower performance compared to the other architectures. It shows a decrease of -16.5% compared to VGG16 ($p = 0.0$), a decrease of -8% compared to NASNetMobile ($p = 0.004$), and a decrease of -19.5% compared to the two MobileNetV3 alternatives ($p = 0.0$).

Finally, regarding the utilization of data augmentation, the statistical analysis verifies that employing this technique does not significantly impact the average performance of the models, obtaining probabilities considerably higher than the significance threshold of 0.05 for all the architectures. Specifically, we observe a probability of $p = 0.625$ for VGG16, $p = 0.857$ for InceptionResNetV2, $p = 0.226$ for NASNetMobile, $p = 0.602$ for MobileNetV3Small, and no difference at all for MobileNetV3Large, obtaining a probability of $p = 1.0$.

We also conducted leave-one-out cross-validation on the best-performing model, namely MobileNetV3Large, for the binary task with and without augmentation. For this experiment, we used the

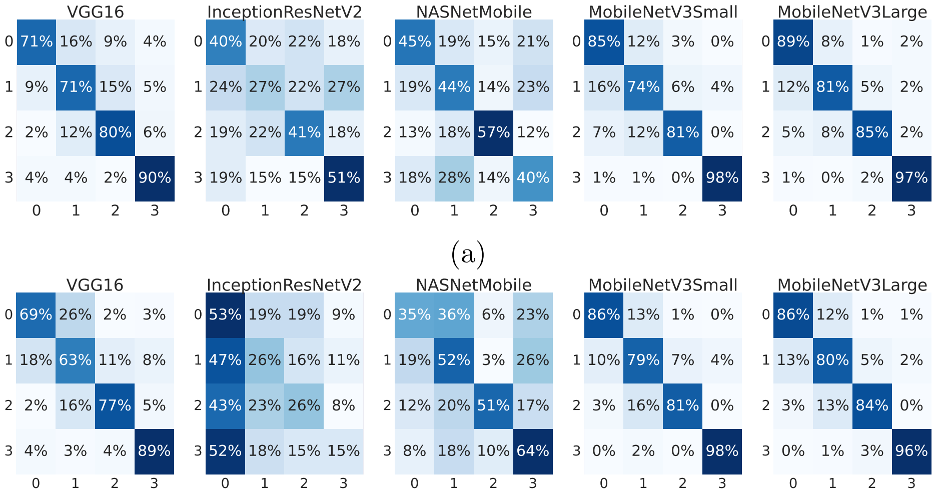
Table 3.2: Multiclass classification performance of the considered base models, with and without data augmentation in the training phase. The performance is reported as mean and standard deviation over the 10-folds of the cross-validation.

Base model	Augmentation	Accuracy	Sensitivity	Specificity	F-1 Score
VGG16	✗	.779 (±.052)	.779 (±.054)	.927 (±.018)	.777 (±.057)
	✓	.745 (±.059)	.744 (±.059)	.915 (±.020)	.738 (±.062)
InceptionResNetV2	✗	.396 (±.087)	.398 (±.088)	.780 (±.023)	.388 (±.084)
	✓	.301 (±.057)	.301 (±.067)	.767 (±.023)	.252 (±.078)
NASNetMobile	✗	.464 (±.073)	.464 (±.073)	.822 (±.025)	.461 (±.076)
	✓	.504 (±.103)	.505 (±.104)	.835 (±.034)	.499 (±.106)
MobileNetV3Small	✗	.846 (±.062)	.847 (±.061)	.948 (±.020)	.843 (±.065)
	✓	.859 (±.054)	.860 (±.052)	.954 (±.017)	.860 (±.049)
MobileNetV3Large	✗	**.882**(±**.057**)	**.881**(±**.055**)	**.960**(±**.019**)	**.879**(±**.058**)
	✓	.866 (±.088)	.866 (±.080)	.956 (±.029)	.863 (±.086)

same hyperparameters as in the best-performing folder after hyperparameter tuning. The results show slightly improved performance (i.e., micro F-1 Score of 0.94 and 0.93 without and with augmentation, respectively) that are due to the larger training set used in this specific evaluation approach.

Multiclass classification task

Table 3.2 summarizes the multiclass classification results of the fine-tuned models, again with and without data augmentation, over the 10-fold cross-validation. It is worth knowing that the specificity in the multiclass setting is the average of the specificity for each class. More specifically, for a given class C, we calculate the specificity of the model based on the one-vs-all approach, thus as the binary problem of distinguishing between samples belonging to C (positive samples) and samples in all other classes (negative samples). Specificity is calculated as true negative, the number of negative cases that are correctly identified as negative, divided by true negatives plus false positives, which is the number of negative cases that are incorrectly identified as positive.



(a)

(b)

Figure 3.6: Confusion matrices related to the multiclass classification setting with original training data (a) and with data augmentation (b). Label 0 refers to Acne samples, label 1 indicates Chickenpox, label 2 indicates *Mpox*, while label 3 indicates the normal class.

Similarly to the binary results, InceptionResNetV2 and NASNetMobile show the worst performances, clearly indicating their inability to recognize the different pathologies in the images. Moreover, data augmentation further reduces the performance of InceptionResnetV2, reducing its F-1 score to 0.252, while it boosts the F-1 score of NASNetMobile to 0.499. In Figure 3.6 we can note in detail how these two models wrongly classify each class and, in particular, how InceptionResNetV2 tends to classify every sample as `Acne` (i.e., class 0). In contrast, VGG16 yields better results, although, similarly to InceptionResnetV2, data augmentation slightly decreases its performance.

The MobileNetV3 variants achieve the best results also in the multiclass setting. MobileNetV3Small yields slightly lower performance: -3.6% in accuracy, -3.4% and -1.2% for sensitivity and specificity, and −3.6% in terms of F-1 score. On the other hand, it benefits more from data augmentation, improving its F-1 score from 0.843 to 0.860. Quite the opposite happens for MobileNetV3Large; in fact, with data augmentation, all its indexes drop. Nevertheless, the confusion matrices clearly show how both of the MobileNetV3 variants are able to successfully identify samples in the `Mpox`, `Acne`, and `Healthy` classes (almost 98% of accuracy, both for augmented and non-augmented models), while `Chickenpox` represents the hardest class, where MobileNetV3Small scores an accuracy of 79% by augmenting the training data, and the larger variant reaches 80% and 81%, respectively with and without data augmentation.

Statistical analysis generally confirms the classification results obtained in our study. Indeed, there were no significant differences found between InceptionResNetV2 and NASNetMobile ($p = 0.188$), which both perform worse than the other considered models. Furthermore, the two variations of MobileNetV3 exhibited a very high probability of $p = 0.776$, suggesting that there were no significant differences between them.

In contrast to the binary classification problem, in the multi-class setting, a noticeable difference can be observed between MobileNetV3Large and VGG16 ($p = 0.0154$), while MobileNetV3Small and VGG16 are similar with a probability of 0.2189. This difference can be attributed to the fact that in the two-sample tests among the three models, the performance of MobileNetV3Small fell between the other two. Indeed, on average, it showed a slight decrease of 3.6% in accuracy compared to its larger variant, while performing better than VGG16 by 6.5%.

Finally, in the case of data augmentation, most of the models did not show statistically significant differences. The probabilities observed were $p = 0.190$ for VGG16, $p = 0.330$ for NASNetMobile, $p = 0.551$ for MobileNetV3Small, and $p = 0.734$ for MobileNetV3Large. Only InceptionResNetV2 showed a probability below the threshold at $p = 0.012$, confirming the largest drop in performance of 6.5% in terms of accuracy.

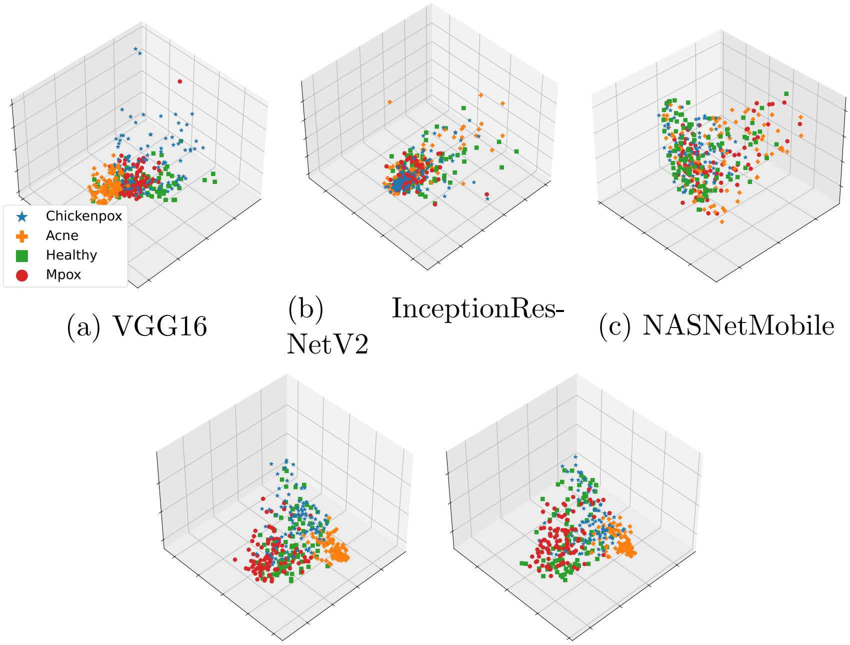
To sum up, we can consider both the MobileNetV3 variants as the best choice to detect mpox from skin lesion images, while the larger model is preferable to accurately distinguish mpox from similar diseases. Moreover, based on the statistical analysis, we can also note that data augmentation does not lead to significant performance improvements, highlighting the need for a larger amount of original training data, as well as a further investigation of more sophisticated approaches of image data augmentation.

Similarly to the binary setting, we conducted a leave-one-out cross-validation for the multiclass classification task. In this case, the results show similar or slightly improved performance (i.e., F-1 Score of 0.90 and 0.85 without and with augmentation, respectively).

Deep embeddings analysis

The obtained results are also supported by the analysis of the deep features (*i.e.*, embeddings) extracted by the different CNNs. Figure 3.7 shows how each model represents the different classes of data samples in the deep latent space, by using Principal Component Analysis (PCA) as data dimensionality algorithm to project the embeddings onto a 3-dimensional plane.

As we can note, for both InceptionResNetV2 and NASNetMobile, it is particularly difficult to distinguish the 4 data clusters: while in the data space modeled by the former CNN, the data points are mainly concentrated in a single blob, in the latter they are distributed on a V-shaped hyperplane, where data of different classes are overlapped to each other. By contrast, the data space modeled by VGG16 makes it easier to distinguish the different classes, even though data points belonging to `Healthy` are still considerably mixed with both `Acne` and `Mpox` samples. The best deep representations are given by the two MobileNetV3 variants, where the considered classes are well-separated. In addition, it is worth noting a lower data dispersion in the MobileNetV3Small embeddings space, thus facilitating the separation of the 4 clusters and, consequently, better classifica-



(a) VGG16 (b) InceptionResNetV2 (c) NASNetMobile

(d) MobileNetV3Small (e) MobileNetV3Large

Figure 3.7: 3-D representation of the dataset based on the deep embeddings learned by each model.

tion performances.

Skin Tone-Based Classification Fairness

It is reasonable to posit that diversity in skin tones may influence the predictive performance of DL models. Consequently, we undertook an additional investigation to assess the models' accuracy in the context of varying skin types.

Since MCSI dataset does not include information regarding the skin tone, we relied on the well-known *Fitzpatrick scale* [195] to classify the available data samples based on the skin pigment. This scale, originally devised within the dermatology field, classifies human skin color into six distinct categories, predicated on the skin's response to ultraviolet (UV) light exposure. The categories range from *Type I*, representing the palest skin that is prone to sunburning and resistant to tanning, to *Type VI*, characterizing deeply pigmented, dark brown skin that does not sunburn easily.

For the purpose of our analysis, we opted to adopt a methodology akin to that employed by Tadesse et al. [196] for categorizing the images into two distinct groups: light and dark skin tones. Specifically, researchers grouped the first four levels of the Fitzpatrick scale under the designation of *Light skin*. Conversely, the fifth and sixth levels were categorized as *Dark skin* tones.

A common approach to annotating images with Fitzpatrick labels is estimating skin tone via *Individual Typology Angle* (*ITA*), which is calculated based on statistical features of image pixels and is negatively correlated with the melanin index [197]. Following the same approach used in [198], we firstly calculated the ITA value of each data sample by using the open-source *Derm-ITA* software[2], and then we mapped values greater than 10 as *Light skin*, while the others as *Dark skin*. At the end of this process, the resulting labels are distributed as follows: `Mpox` 57 Light and 43 Dark; `Chickenpox`, 78 Light and 22 Dark; `Acne`, 73 Light and 27 Dark; and finally, `Healthy` 69 Light and 31 Dark.

Based on this distinction between light and dark skin, we evaluated the models' performance (without retraining the models) in both binary and multiclass scenarios, accounting for the two distinct skin types. The summarized results are presented in Table 3.3, showing the average accuracy values and their corresponding standard deviations.

The statistical analysis (i.e., standard t-test) highlights some significant differences only in the binary classification task, showing better performance in classifying the under-represented class, that is, dark skin samples. Specifically, in the binary classification setting, MobileNetV3Large without data augmentation obtains significance

[2] `https://github.com/AdamCorbinFAUPhD/derm_ita/tree/master`

Table 3.3: Average classification accuracy (and standard deviation) for the two types of skin tones in binary and multiclass settings.

		Binary		Multiclass	
Base model	Augmentation	Light	Dark	Light	Dark
VGG16	✗	.793 (±.118)	.886 (±.100)	.774 (±.058)	.766 (±.158)
	✓	.774 (±.052)	.899 (±.054)	.733 (±.062)	.757 (±.163)
InceptionResNetV2	✗	.586 (±.071)	.551 (±.096)	.413 (±.081)	.321 (±.149)
	✓	.556 (±.114)	.521 (±.106)	.307 (±.071)	.276 (±.118)
NASNetMobile	✗	.673 (±.088)	.785 (±.096)	.474 (±.081)	.428 (±.147)
	✓	.676 (±.105)	.786 (±.086)	.496 (±.085)	.481 (±.159)
MobileNetV3Small	✗	.839 (±.093)	.921 (±.077)	.854 (±.093)	.820 (±.131)
	✓	.851 (±.089)	.883 (±.072)	.850 (±.063)	.857 (±.111)
MobileNetV3Large	✗	.773 (±.106)	.965 (±.059)	.876 (±.061)	.868 (±.086)
	✓	.835 (±.101)	.919 (±.077)	.850 (±.113)	.862 (±.143)

of $p = 0.000881$, while VGG16 and NASNetMobile with data augmentation show significance values of $p = 0.000092$ and $p = 0.025547$, respectively. One plausible explanation for this phenomenon could be the higher contrast between skin tone and skin lesion colors in the case of dark skin samples. This contrast likely aids the DL models in accurately identifying conditions such as mpox and the other considered pathologies from skin images.

3.3 Analysis of Grad-CAM indications

Gaining a more profound comprehension of deep learning models, often perceived as "black-boxes", is important in the context of medical applications. Specifically, the field of Explainable Artificial Intelligence (XAI) has emerged with dual objectives: enhancing model interpretation and allowing additional validations of the model results.

One notable XAI technique, Grad-CAM, assumes significance in this pursuit by enabling the identification of salient features that drive the model's predictions. Consequently, it serves as a valuable adjunct tool for delving into the rationale underpinning the decisions made by the model.

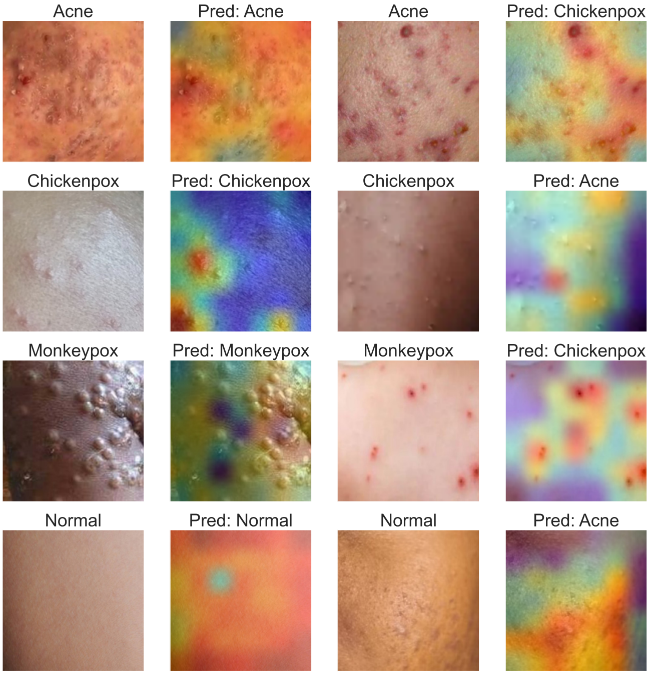


Figure 3.8: Examples of Grad-CAM results for each class with MobileNetV3Large, first and third columns show the input image, (Correctly and wrongly predicted respectively). Second and fourth columns show Grad-CAM explanations (for correctly and misclassified examples)

We decided to apply Grad-CAM to the predictions provided by MobileNetV3Large as one of the best-performing models in both classification tasks. Specifically, in order to understand what features of the input images are considered relevant by the model, in Figure. 3.8 we reported 8 different examples of explanations, four correctly predicted, along with their class activation maps (first and second columns), and four misclassified samples, with their corresponding maps (third and fourth columns). The ground-truth label and the predicted one are indicated at the top of each image, while the heatmaps have been generated by superimposing the class activation map to the original image. While bluish areas identify less relevant features for the given class, warmer colors (e.g., orange and red) represent the most relevant ones that have led the models to provide the specified prediction.

For example, the first row represents a case of `Acne`. When the model correctly classifies the image, the relevant features are distributed across all scars and pustules, which are typical of a strong presence of acne. However, when the model misclassifies the image, the main focus of the network is on the pimples, neglecting the skin scars, causing the model to classify the image as `Chickenpox`.

Regarding the `Chickenpox` sample, when the model provides a correct prediction, its focus is only on the largest pimples, whereas when the model makes an incorrect prediction, its attention is distributed to minor skin defects in addition to the pimples, classifying the image as `Acne`.

For `Mpox`, the model is capable of correctly identifying the pathology when vesicles and crusts are formed, but it clearly fails in the early stages of the pathology, when pimples have not yet fully developed, providing a wrong prediction (i.e., `Chickenpox` in this case).

Finally, when the model correctly classifies a `Healthy` image, as we can expect, the importance of the feature is evenly distributed throughout the image without focusing on specific elements. On the contrary, when the model misclassifies a healthy sample, it is because it gives great relevance to hair and skin damage, classifying the image as `Acne`.

The model's visual attention analysis shows that MobileNetV3Large effectively identifies reasonable features for each class. The model's misclassifications are justifiable due to the similarity of the different classes, and, despite these errors, the model's overall ability to identify relevant features highlights its potential in our specific use-case scenario, providing more reliability on the model's predictions.

Table 3.4: Model sizes and classification performance with mobile optimization.

Task	Base model	Quant.	Size (MB)	Accuracy	Sensitivity	Specificity	F-1 Score
binary	VGG16	✗	268.44	.894 (±.049)	.841 (±.075)	.740 (±.143)	.851 (±.070)
		✓	67.22	.894 (±.053)	.841 (±.077)	.740 (±.143)	.851 (±.072)
	InceptionResNetV2	✗	350.53	.735 (±.056)	.562 (±.110)	.220 (±.266)	.533 (±.130)
		✓	89.42	.702 (±.113)	.585 (±.105)	.350 (±.310)	.548 (±.127)
	NASNetMobile	✗	336.37	.830 (±.036)	.765 (±.032)	.640 (±.070)	.769 (±.036)
		✓	85.03	.738 (±.060)	.750 (±.065)	.780 (±.123)	.702 (±.058)
	MobileNetV3Small	✗	211.05	.932 (±.043)	.883 (±.070)	.790 (±.129)	.902 (±.064)
		✓	53.01	.915 (±.046)	.851 (±.067)	.730 (±.116)	.875 (±.068)
	MobileNetV3Large	✗	382.93	.928 (±.042)	.884 (±.090)	.800 (±.200)	.891 (±.074)
		✓	96.17	.923 (±.051)	.875 (±.106)	.780 (±.225)	.884 (±.089)
multiclass	VGG16	✗	318.21	.779 (±.053)	.779 (±.054)	.927 (±.018)	.777 (±.057)
		✓	79.63	.782 (±.044)	.782 (±.044)	.927 (±.015)	.779 (±.050)
	InceptionResNetV2	✗	485.12	.398 (±.088)	.398 (±.088)	.799 (±.027)	.388 (±.084)
		✓	122.48	.308 (±.064)	.306 (±.065)	.769 (±.022)	.243 (±.075)
	NASNetMobile	✗	259.23	.470 (±.074)	.470 (±.074)	.822 (±.026)	.467 (±.076)
		✓	65.51	.471 (±.095)	.471 (±.095)	.823 (±.034)	.449 (±.102)
	MobileNetV3Small	✗	225.77	.847 (±.061)	.847 (±.055)	.949 (±.014)	.843 (±.065)
		✓	55.62	.833 (±.066)	.833 (±.066)	.944 (±.020)	.831 (±.066)
	MobileNetV3Large	✗	278.73	.881 (±.055)	.881 (±.055)	.962 (±.018)	.879 (±.058)
		✓	69.98	.880 (±.046)	.879 (±.046)	.961 (±.014)	.875 (±.052)

3.4 Mobile optimization

Table 3.4 shows the great advantage of using quantization to reduce the memory footprint of the models without requiring their retraining. As we can note, the original size of the DL models trained for mpox detection considerably varies for the different base architectures, ranging between 200 MB and almost 500 MB, which can limit their implementation on several personal mobile devices. On the other hand, by using quantization to lower the operations' precision from 32-bit floats to 16-bit floats, all the models' sizes are reduced by approximately 4 times. For example, the size of VGG16 tuned for binary classification dropped from 268.44 MB to just 67.22 MB, while the size of InceptionResNetV2 for multiple classes (i.e., the most demanding model in terms of memory) has been reduced by 74.75%, limiting its memory footprint from 485.12 MB to 122.48 MB.

Furthermore, it is important to highlight that the impact of quantization on the classification performance of the majority of the examined architectures remains relatively modest, resulting in an average reduction of no more than 1% in accuracy.

However, it is noteworthy that InceptionResNetV2 and NASNetMobile exhibit more pronounced performance penalties due to quantization. Specifically, InceptionResNetV2 experiences a decline of approximately 3% and 9% in accuracy in the binary and multiclass settings, respectively. Meanwhile, NASNetMobile's accuracy registers a noteworthy 10% reduction, albeit exclusively in the binary task. Remarkably, in the multiclass experiments, it performs nearly on par with its non-quantized counterpart. We suspect that this can be attributed to the inherent effect of quantization, which compromises the precision of both weight parameters and activation functions. Consequently, this effect is more pronounced in larger networks, such as InceptionResNetV2 and NASNetMobile. Additionally, it is worth noting that these models already exhibit relatively lower accuracy levels prior to quantization, and when this factor is coupled with quantization, it results in more substantial performance losses compared to the other models.

Besides the memory size and classification performance, we also conduct an empirical evaluation of the models' time complexity. Even though our application scenario does not require real-time predictions, fast computation represents a key requirement when dealing with mobile personal devices like smartphones. Therefore, to perform this type of experiment, we rely on the benchmark tool provided by TensorFlow Lite (TFLite) [3], the Google-released mobile library for deploying models on mobile devices, microcontrollers, and other edge devices. Specifically, we first convert our CNN models to the TFLite format; then, we deploy such models on the TFLite Android benchmark app[4] that executes each model 50 times with synthetic input to collect reliable statistics related to the inference times on a real An-

[3] `https://www.tensorflow.org/lite`

[4] `https://www.tensorflow.org/lite/performance/measurement`

Table 3.5: Average inference times (in seconds) on different mobile devices, by using both CPU (4 threads) and GPU for the computation.

			Google Pixel 6a		Xiaomi Mi 9T	
Task	Base model	Quant.	CPU	GPU	CPU	GPU
binary	VGG16	✗	.429 (±.051)	.031 (±.002)	.606 (±.013)	.245 (±.011)
		✓	.104 (±.013)	.031 (±.002)	.430 (±.021)	.245 (±.011)
	InceptionResNetV2	✗	.134 (±.012)	.057 (±.007)	.515 (±.050)	.188 (±.016)
		✓	.064 (±.005)	.057 (±.007)	.441 (±.039)	.188 (±.016)
	NASNetMobile	✗	.041 (±.014)	.023 (±.003)	.206 (±.051)	.062 (±.029)
		✓	.033 (±.005)	.023 (±.003)	.421 (±.037)	.060 (±.027)
	MobileNetV3Small	✗	.018 (±.004)	**.011** (±**.002**)	.056 (±.014)	.033 (±.010)
		✓	**.011** (±**.002**)	**.011** (±**.002**)	.104 (±.023)	**.032** (±**.010**)
	MobileNetV3Large	✗	.018 (±.010)	.013 (±.004)	.067 (±.028)	**.032** (±**.019**)
		✓	.014 (±.004)	.013 (±.003)	.140 (±.040)	**.032** (±**.020**)
multiclass	VGG16	✗	.423 (±.067)	.031 (±.002)	.612 (±.012)	.249 (±.012)
		✓	.117 (±.084)	.031 (±.002)	.196 (±.007)	.249 (±.012)
	InceptionResNetV2	✗	.139 (±.008)	.059 (±.008)	.243 (±.007)	.192 (±.020)
		✓	.065 (±.004)	.059 (±.008)	.141 (±.016)	.192 (±.020)
	NASNetMobile	✗	.036 (±.009)	.021 (±.002)	.084 (±.022)	.051 (±.021)
		✓	.031 (±.003)	.021 (±.003)	.127 (±.015)	.052 (±.021)
	MobileNetV3Small	✗	.011 (±.007)	.009 (±.003)	.028 (±.016)	**.024** (±**.015**)
		✓	**.008** (±**.003**)	.009 (±.003)	.040 (±.009)	.040 (±.009)
	MobileNetV3Large	✗	.016 (±.005)	.012 (±.001)	.047 (±.014)	.029 (±.007)
		✓	.014 (±.002)	.012 (±.001)	.062 (±.004)	.029 (±.007)

droid smartphone. Moreover, in order to get insights on the models' performance on different hardware settings, we perform our evaluation on 2 smartphones, by using both CPU (with multithreading) and GPU for the computation: (i) a recent Google Pixel 6a released in 2022, with the latest Android 13 operating system, an Octa-Core CPU (2x2.80 GHz Cortex-X1, 2x2.25 GHz Cortex-A76, and 4x1.80 GHz Cortex-A55), and the Mali-G78 MP20 GPU; and (ii) an older Xiaomi Mi 9T, released in 2019, with Android 10, an Octa-core CPU (2x2.2 GHz Kryo 470 Gold and 6x1.8 GHz Kryo 470 Silver), and an Adreno 618 GPU.

Table 3.5 summarizes the average inference times (in seconds) of the considered models in the different hardware settings, both for the binary and multiclass classification tasks, highlighting in bold face the best results for each device and task. It is clear that even the largest models such as VGG16 and InceptionResNetV2 can provide a prediction in less than 0.612 seconds when deployed on modern smartphones. The benefit of using quantization can be mainly observed when the computation is based on CPU, reducing the inference time by 50% at least in some cases (e.g., VGG16 and InceptionResNetV2 with Google Pixel 6a). On the other hand, all models can be executed by the GPU in less than 0.059 seconds on Google Pixel 6a and 0.245 seconds on Xiaomi Mi 9T, thanks to its ability to parallelize all operations that are involved in a deep neural network [199].

Finally, we can also note that the CNN that performs best in terms of classification accuracy, i.e., MobileNetV3 (both Small and Large variants), is also the one with the lowest inference time. In fact, while the larger variant provides a prediction for binary and multiclass classification, respectively, in not more than 0.018 and 0.016 seconds on Google Pixel 6a and not more than 0.140 and 0.062 seconds with Xiaomi Mi 9T, MobileNetV3Small requires only not more than 0.018 and 0.011 seconds on the Google phone and not more than 0.104 and 0.040 seconds on the Xiaomi, thus proving the feasibility of efficiently performing the whole data processing and prediction tasks directly on mobile devices.

3.5 Conclusion

The Chapter introduces a novel m-health system for the preliminary screening of mpox infections through pictures of skin rashes and eruptions taken with common smartphone cameras. The system is designed to be entirely executed on mobile devices and is characterized by the use of Transfer Learning to adapt state-of-the-art Convolutional Neural Network (CNN) models for image classification, mobile-oriented optimization of the models through quantization, and the use of Grad-CAM as eXplainable AI (XAI) technique for technical validation.

While the proposed solution cannot replace the expertise of a medical professional, it serves as a preliminary alert system for self-examination in at-home settings, particularly in areas with limited medical assistance and where continuous Internet connectivity is not assured. In addition, such a system can play a pivotal role for supporting the preliminary screening of large populations, alleviating the burden on medical facilities, and limiting the dissemination of the virus, aiding in the prompt identification of emerging outbreaks by detecting new cases as soon as they arise.

The models have also been evaluated for their complexity in terms of execution time on commercial smartphones, and they all obtained performances under 1 second to provide the prediction, with quantization further reducing the inference time on CPUs.

Despite achieving promising results, our study has four main limitations. First, the limited number of training data. Second, the lack of other metadata information, that can help evaluate the data heterogeneity with respect to various factors like gender, race, age, and physical conditions. This is clearly relevant for ethical data collection and fair model training. Third, MCSI contains images derived from online resources that were manually selected and cropped by a skilled operator, while in the intended application the images will be self-acquired and possibly cropped by the end-user or a caregiver by following the application instructions. We cannot exclude that self-acquired images will have different properties that can impact the performance of the detection models. The fourth limitation is related to annotations' reliability, in terms of skin lesion type: MCSI derives the annotations from the existing datasets and the source of the annotations is not specified.

A possible solution to address the first three problems above is to release a prototype application implementing the proposed detection system. The application could help remotely collect new images, hence creating a larger dataset to improve the current detection model. Also, the application could easily collect additional user information, like gender and age. Another advantage of this solution is that the images would be collected by the end-users or their caregiver. In order to address the fourth limitation, but also to effectively design the proposed application and clinically validate the related results, it is essential to establish a strict collaboration with medical experts, especially dermatologists and virologists. The collaboration could also provide additional data to further investigate the algorithms performances.

4

Multi-Task Learning for Ultrasound Detection of Subquadricipital Recess Distension

The previous chapter described a complete framework to detect MPOX from mobile phone images adopting Transfer Learning to deal with the scarce available data. This showed promising results, but, as we will discuss in the following, Transfer Learning alone might not be a sufficient solution to achieve acceptable performance on a more complicated task. Despite its limitations, Transfer Learning serves as a critical foundation for this thesis, providing the essential starting point upon which more advanced methods are built and refined. Here, we formulate the research problem of supporting physicians in diagnosing joint recess distension in patients with hemophilia using a CAD system. The problem consists of detecting the joint recess within US images and classifying it as *Distended* or *Non-distended*. Specifically, we focus on the main joint recess of the knee, also called *SubQuadricipital Recess* (SQR). We consider the SQR longitudinal scan, which is one of the three scans specified in the HEAD-US protocol for this joint [14].

In this chapter, we propose two approaches to address the distension detection problem formulated in Section 1.2. The first one, called the *Detection approach*, adopts state-of-the-art object detection to find *Distended* or *Non-distended* SQR inside the US image and returns the detection having the highest confidence. The second solution, called the *Multi-task approach* uses a multi-task learning process, with the aim of simultaneously detecting the SQR inside the US image and classifying it as *Distended* or *Non-distended*.

The experiments were conducted on the SQR knee dataset described in Section 2.7. The experiments, we compared the two proposed solutions among themselves and with two baselines based on transfer learning, one *Classification baseline* and one *Detection baseline*. The results reveal that both the *Multi-task approach* and the *Detection approach* improve over the *Classification baseline* in terms of balanced accuracy. Furthermore, the *Multi-task approach* outperforms both the *Classification baseline* and the *Detection approach* in terms of balanced accuracy and sensitivity, which, as we motivate in the following, is particularly relevant for the given problem. Concerning detection accuracy, the *Detection approach* has a slightly better performance than the *Multi-task approach*, and remains in line with the *Detection baseline*.

4.1 Problem modeling

An interview, conducted with physicians from the Angelo Bianchi Bonomi Hemophilia and Thrombosis Center (two of which are also authors of this work), revealed the need for a computer aided tool (CAD) supporting the physician in diagnosing SQR distension. The tool can be used as a part of a protocol for the early diagnosis of hemarthrosis, which is particularly relevant for hemophilic patients [16, 11]. Indeed, directly identifying hemarthrosis in US images is particularly challenging as it requires to distinguish blood from synovial fluid and blood clots from synovial hyperplasia, which appears very similar.

To support the physician during the diagnosis, the CAD tool should identify the position of the SQR inside the specified US scan and classify it as *Distended* or *Non-distended*.

In terms of machine learning, the CAD tool needs to implement a combination of classification and detection techniques. For what concerns the classification, existing models can be directly applied to the given problem, defining two classes, one for the *Distended* and the other for the *Non-distended* recess.

For what concerns the detection problem, we model the recess as the target object to detect. Two possible solutions can be adopted: to model two distinct classes of objects (*i.e.*, one for *Distended* and another for *Non-distended* recesses) or to model a single class (*i.e.*, representing both *Distended* and *Non-distended* recesses). In both cases, the direct application of existing object detection algorithms would not correctly model the given problem. Indeed, most of the existing object detection techniques assume that multiple objects can be detected in a single image, from the same or different classes. This is appropriate, for example, in the problem of tumor detection, since multiple malign and benign tumors can be visible in the same image [200]. Instead, in the given problem, we can infer from domain knowledge that a single object (*i.e.*, a recess) is visible in each image.

As we show in the following, with the *Detection approach* we model two distinct classes, while with the *Multi-task approach* we model a single class. Also, both solutions extend existing object detection techniques by returning a single object for each input image.

4.2 Methodology

We propose two solutions for the problem defined in Section 1.2. The first solution, which we name *Detection approach*, is described in Section 4.2.1. It is based on a state-of-the-art detection technique, adapted to solve both the detection and the classification problems. The second solution, which we call *Multi-task approach* (see Section 4.2.2), is a multi-task network with a branch that solves the detection problem and another one that solves the classification problem.

4.2.1 *Detection approach*

Figure 4.1 depicts the network architecture of the *Detection approach*. Each input US image is processed by the YoloV5 [201] object detec-

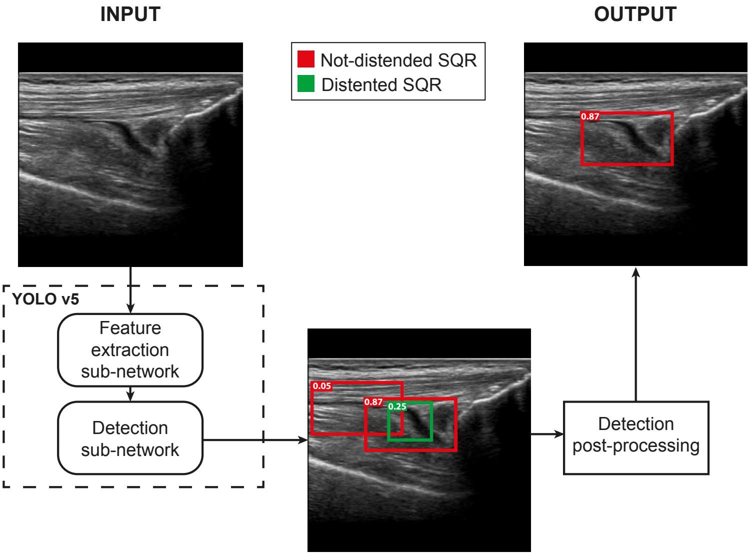


Figure 4.1: Overall architecture of the *Detection approach*

tor that returns a set of candidate SQRs, each characterized by a confidence value, a bounding box and the label (*Distended* or *Non-distended*). Since in the considered domain, the input image actually contains exactly one SQR, the *Detection Post-processing* module selects the prediction with the highest confidence and outputs its bounding box and its label.

We train the network to recognize two classes of objects: *Distended* SQRs and *Non-distended* SQRs. Since the amount of labeled images in this domain is generally scarce, it is difficult to collect a sufficiently large dataset to fully train a robust detection network. Therefore, we adopt a transfer learning approach [64] to initialize the network's weights. Specifically, we use the pre-trained weights publicly available for the *YoloV5* network, trained on the MS COCO dataset [81]. Finaly, the network is fine-tuned on the US images by freezing the encoder, and keeping the sub-network trainable.

YoloV5 is a single-stage detector designed to detect different objects in an image and directly assign them the corresponding class. *YoloV5* is an optimized version of the *YoloV4* framework [202], that has been widely used in the literature for object detection tasks. Specifically, among the five models available in *YoloV5*, we use the *large* model, which was empirically selected as it achieved the best results in preliminary tests. *YoloV5* is internally divided into a feature extraction sub-network and a detection sub-network. It also adopts a specific loss function and an early stop criterion. These four concepts are briefly described in the following.

Feature Extraction sub-network The *Feature Extraction sub-network* is a Convolutional Neural Network (CNN). Specifically, it is a *CSPDarknet53* network, that was originally proposed in [203] and that was shown to be particularly effective for object detection [202] and US image classification [204].

Detection sub-network The *Detection sub-network* is divided into a *neck* and a *head* parts.

The overall goal of the *neck* part is to divide the image into multiple small fragments with the objective of simplifying further analysis by performing semantic segmentation (by associating categories to pixels) as well as instance segmentation (classifying and locating objects at pixel level). The *head* part is a one-stage detector [205] that processes the features returned by the *neck* part and outputs the bounding boxes of the detected elements along with their predicted class.

Loss function We use the default *YOLOV5* loss function that is shown in Equation 4.1 and that is computed as the weighted sum of three values: a) the *localization loss* (L_{box}) is computed with the *Complete IoU* loss function (CIoU) [206], and represents the error in the position of the predicted bounding box; b) the *class loss* (L_c) is computed with Binary Cross-Entropy (BCE) and represents the error in classifying the predicted class; c) the *objectness loss* (L_{obj}) is computed with BCE and represents to which extent the predicted bounding box actually encloses an object of interest. The weights of these values are hyper-parameters that need to be empirically tuned (see Section 4.3.4).

$$L = \alpha L_{box} + \beta L_{obj} + \gamma L_c \tag{4.1}$$

Early stopping criterion We use the default *YOLOV5* early stopping criterion to terminate the training if there are no improvements in the results for a given number of training epochs. This default criterion considers the mean Average Precision (mAP) of the detection, *i.e.*, the ratio of correctly classified bounding boxes considering a given threshold of the IoU with the corresponding ground truth. Note that, in a multi-class scenario, this criterion factors for both the correct classification and the correct detection of the objects. Specifically, it is computed as the weighted sum of the mAP@0.5 and the mAP@0.5:0.95 where a weight of 0.1 is given for mAP@0.5, and a weight of 0.9 is given for mAP@0.5:0.95 in order to prioritize more accurate bounding boxes detection.

4.2.2 *Multi-task approach*

The *Detection approach* addresses the problem of classifying the SQR as *Distended* or not, by selecting the label of the detection with the highest confidence. An alternative (and possibly more natural) solution would be to classify the entire image. However, this would not provide the needed SQR bounding box. For this reason, we propose the *Multi-task approach* that pairs image classification and detection.

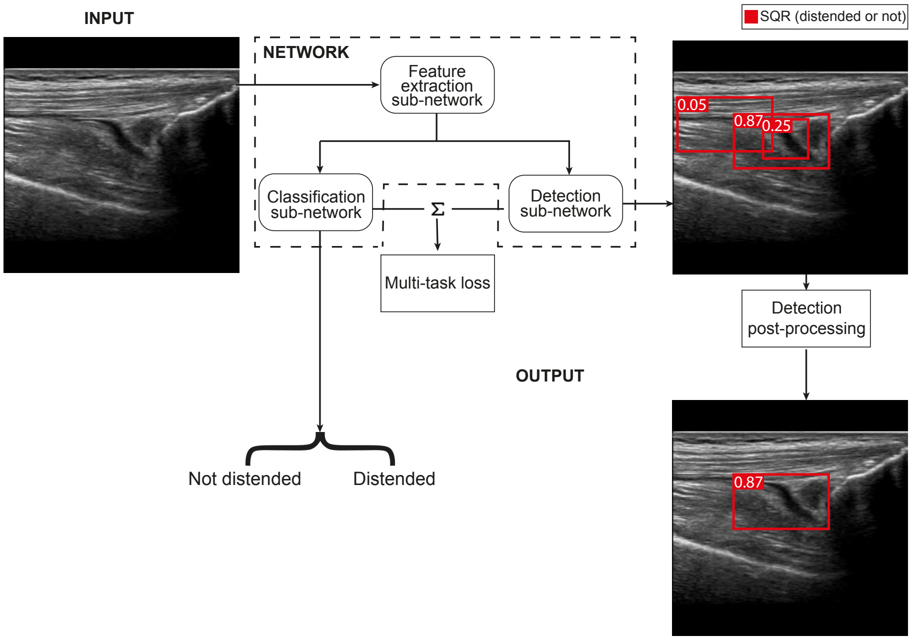


Figure 4.2: Overall architecture of the *Multi-task approach*

The proposed network is a modified version of the network used for the *Detection approach*. The key modification consists of a *Classification sub-network* that performs the SQR binary classification. The input image is first processed by the *Feature Extraction* sub-network, that is shared for both classification and detection tasks. Then the extracted features are simultaneously processed by the *Detection sub-network* and the *Classification sub-network*. The *Classification sub-network* processes the features and returns the predicted SQR class (*i.e.*, distended or not) considering the whole image.

Differently from the *Detection approach* solution, the goal of the *Detection sub-network* in the *Multi-task* solution is simply to detect the SQR, without providing information about the distension. Hence, the *Detection sub-network* network is trained with a single class and it returns a set of bounding boxes, all belonging to the same class, each with an associated confidence value. The *Detection Post-processing* module selects the bounding box with the highest confidence. During the training phase, the *multi-task loss* jointly considers the errors on classification and detection to update the network weights.

Classification sub-network

Figure 4.3 shows the *Classification sub-network* of the *Multi-task Approach*. The first layer of the sub-network is an Adaptive Average Pooling Layer in charge of reducing the feature dimensions to a fixed 2-dimensional output size. Then, the output is provided to a Flatten Layer, that converts 2-dimensional data to a 1-dimensional array. This array is then processed by a fully connected network composed of two hidden layers of 1024 and 512 units, respectively. These layers use a *ReLu* activation function. A dropout layer is applied between the two hidden layers with the objective of reducing overfitting. Finally, a Softmax layer is in charge of providing the most likely class (*i.e.*, *Distended/Non-distended*). The architecture of this network has been determined empirically, during the tuning phase the network is kept trainable and the layers of the classification head are initialized with a normal distribution centered on 0.

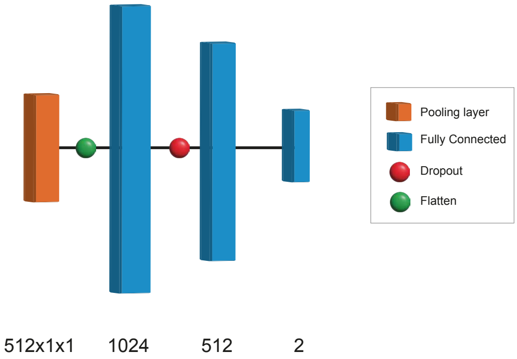


Figure 4.3: Classification sub-network architecture

Multi-task loss

Training the multi-task network requires a custom loss function that simultaneously takes into account the classification and detection errors. For this reason, we adapt the loss function used for the *Detection approach* by adding a new loss term that represents the errors of the *Classification sub-network*. Specifically, we adopt a typical solution in binary classification that consists in computing the classification error L_{cls} with a BCE function. Another difference with respect to the loss function used in the *Detection approach*, is that, in the *Multi-Task approach*, the *Detection sub-network* is trained with a single class, hence there are no possible errors with class prediction. Thus, the L_c parameter, considered in Equation 4.1, is always zero. So, the overall multi-task loss is computed as the weighted sum of L_{box}, L_{obj}, and L_{cls}, as shown in Equation 4.2. These weights are hyper-parameters that need to be empirically tuned (see Section 4.3.4).

$$L = \alpha L_{box} + \beta L_{obj} + \delta L_{cls} \tag{4.2}$$

Since the datasets in this domain are usually highly unbalanced (*e.g.*, in our dataset $\approx 75\%$ of the images are labeled as *Non-distended*), there is the risk that the network favors *Non-distended* classifications, which in turn may increase the number of false negatives. In order to mitigate this problem, we adjust the classification loss L_{cls} to give higher error values to false negatives (*i.e.*, *Distended* SQR classified as *Non-distended*). This is achieved by adding an additional weight to L_{cls} when the ground truth is *Distended*. Specifically, to achieve a balanced classification, the weight is computed as the ratio between the *Non-distended* and *Distended* samples in the training set. Thanks to this approach, the errors on the *Distended* samples have a more significant impact on the overall loss.

Multi-task early stopping criterion

As specified above, for the *Detection approach*, the default *YOLOV5* early stopping criterion, based on mAP, is used to stop the training if no improvements are detected for a specified number of epochs.

Instead, for the *Multi-task approach*, since the detection is computed for a single class, the mAP does not account for the classification accuracy but only considers the detection accuracy. Thus, for the *Multi-task approach*, we consider a weighted sum of mAP@0.5 for the detection and balanced Accuracy for the classification on the validation set. In particular, we provide a higher weight (0.7) to the balanced accuracy and a lower one to mAP@0.5 (0.3). This is due to the fact that we prefer to be more accurate on the classification, at the cost of identifying slightly less accurate (but still informative) bounding boxes. We consider a patience value of 100 epochs, which means that the training is stopped if the early stopping criterion does not improve for the number of epochs specified by the patience value.

4.3 Evaluation

In this section, we describe the experimental evaluation conducted on the dataset introduced above. First, we present the baselines used in the study. Then, we describe the adopted evaluation methodology, the metrics and we describe how we selected the hyper-parameters. Finally, we show the results of the two proposed solutions and compare them among themselves and with the two baselines. We conclude the section by showing examples of the application of the proposed solutions and by discussing the results.

4.3.1 Baselines

To evaluate the effectiveness of the two proposed solutions, we compared them with two baselines, one for each of the two tasks that we address: classification and detection.

The *Classification baseline* is a binary classifier that uses *Darknet53* [205] as feature extractor (*i.e.*, the same one as in the *Multi-Task* and *Detection* approaches). The feature vector is then passed to a fully connected layer that performs the classification. As in our proposed solutions, the feature extractor was pre-trained and frozen during training. We consider this approach as a baseline for the classification recognition rate since it represents a widely adopted solution for medical image classification [207].

The *Detection baseline* is a object detector with the same architecture as the *Detection approach*. The main difference with respect to the *Detection approach* is that the *Detection baseline* detects a single class, the SQR, without considering whether it is distended or not. The *Detection baseline* outputs the object detected with the highest confidence. We selected this solution as a baseline for the detection task because the technique is widely adopted in the literature [207] and, differently from the *Detection approach*, it only focuses on the SQR detection task without considering the classification task. Since the *Detection baseline* addresses a simpler problem than our solutions, it represents an upper bound for the detection performance of our solutions.

In order to fairly compare the four techniques (two baselines and the two proposed solutions), the data follows the same pre-processing and training pipelines described in Section 4.3.3. For the same reason, all four techniques are evaluated using the same cross-validations splits.

4.3.2 Metrics

We define two sets of metrics: one for the detection and the other for the classification. For what concerns the detection, we measure the average Intersection over Union (IoU). The IoU between two plane figures is defined as the ratio between the area of their intersection and the area of their union. When measuring the performance of a given technique, for each test image we measure the IoU between the predicted bounding box and the ground truth bounding box. Then, we compute the average of this metric among all test images. Prior literature commonly considers as correct the detections with an IoU $\geq$ than 0.5 [208]. Thus, we consider this as a threshold for an acceptable IoU result.

Considering classification, for each image we compare the ground truth class with the predicted class hence computing if the result is a True Positive (TP), a True Negative (TN), a False Positive (FP), or a False Negative (FN). Note that the positive class is *Distended* and the negative class is *Non-distended*. Then, we used the following classification metrics:

- Specificity: measures the ability of the model to identify true negatives. Specificity is defined as $\frac{TN}{TN+FP}$
- Sensitivity: measures the ability of the model to identify true positives. Sensitivity is defined as $\frac{TP}{TP+FN}$
- Balanced accuracy: mean between specificity and sensitivity. It is considered a sounder metric compared to accuracy when the class imbalance is high [209]. Balanced accuracy is defined as $\frac{sens+spec}{2}$
- Confidence interval (CI): the 95% confidence interval for the classification and detection results. The CI provides a reliability measure of the results by indicating the range in which the results of the repetitions of the same experiment should fall 95% of the time, thus showing the consistency level of the reported results [210].

4.3.3 Evaluation methodology

The evaluation of the recognition rate of the proposed solutions is based on a 5-fold cross-validation. In order to avoid high correlation bias, the training and the test splits do not have images from the same patients in common. The consequence is that we could not exactly divide the dataset in 80% and 20% splits and therefore the splits have a slightly different number of images.

An example fold subdivision can be found in Table 4.1. Each training fold was further split: 80% as training set and 20% as validation set. During training we used SGD with momentum [211] as optimizer.

Fold 0	**Train**	**Test**	**Total**
Non-distended	289	71	360
Distended	97	26	123
Total	386	97	483
Total patients	166	42	208

Table 4.1: Example data distribution in Fold 0 of the 5-fold cross-validation

4.3.4 Hyper-parameters selection

In order to properly tune the many hyper-parameters of our network, we adopt an evolutionary approach [212]. Given a fitness function, an evolutionary algorithm evaluates the best fitting set of hyper-parameters thanks to *mutation* and *cross-over* operations. For the sake of this work, we considered the evolutionary method proposed in *YOLOV5*, that only considers the mutation operation with 90% of probability and 0.04 of variance. Each mutation step generates a new set of hyper-parameters given a combination of the best parents from all the previous generations. The fitness functions used for the hyper-parameters selection for the *Detection approach* and the *Multi-task approach* correspond to the early stopping criteria introduced in Sections 4.2.1 and 4.2.2, respectively.

In order to balance the need for a high number of evolution epochs with limited computational resources, we run the evolutionary algorithm only on one fold. We executed our evolutionary algorithm for 300 epochs on each solution. Considering the *Multi-task approach*, the best results have been obtained at the 193th epoch, while for the *Detection approach* the best set of hyper-parameters was found at the 4th epoch. The set of hyper-parameters resulting from evolution has been used to evaluate our approaches on the complete cross validation procedure. The most relevant discovered hyper-parameters are presented in Table 4.2

Note that γ is a weight associated to the L_c loss that is only

	Learning rate	**Dropout**	**SGD momentum**	α	β	γ	δ
Detection	0.00369	-	0.77628	0.06868	0.49062	0.2343	-
Multi-task	0.0018	0.11008	0.62403	0.05427	0.67598	-	0.41855

Table 4.2: Selected hyper-parameters

considered in the *Detection approach*, while δ is a weight associated to the L_{cls} loss that is only considered in *Multi-task approach.* Finally, the Dropout rate is only included in the *Classification sub-network* of the *Multi-task approach.*

4.3.5 Results

Table 4.3 shows the performance of the two baselines and of the two proposed solutions. Note that, in order to fairly compare the *Detection approach* with the *Detection baseline* and the *Multi-task approach*, the average IoU for the *Detection approach* (marked with *) is computed ignoring the predicted class. This means that, for the detection approach, we consider the bounding-box of the detection with the highest confidence, without considering if the class of the detected box is actually correct.

	Balanced accuracy	**Specificity**	**Sensitivity**	IoU
Classification baseline	0.73 ± 0.03 [0.72 - 0.74]	0.85 ±0.09	0.61 ± 0.13	-
Detection baseline	-	-	-	**0.66 ± 0.02**
Detection Approach	0.74 ± 0.07 [0.73 - 0.75]	**0.97 ±0.03**	0.52 ± 0.12	**0.66 ± 0.01*** [0.65 - 0.66]
Multi-task Approach	**0.78 ± 0.05** [0.77 - 0.79]	0.92 ± 0.04	**0.64 ± 0.09**	0.63 ± 0.02 [0.62 - 0.63]

Table 4.3: Evaluation results (reported as mean among the folds ± standard deviation), 95% CI are reported between []

Since both the early stopping criterion and the hyper-parameters selection methods for the *Multi-task approach* are designed to prioritize the classification accuracy at the expense of the detection accuracy, its balanced accuracy is confirmed to be higher than for the *Detection approach.* Specifically, the *Detection approach* has a balanced accuracy of 0.74 (95% CI $[0.73 - 0.75]$), slightly improving over the *Classification baseline* which reaches a balanced accuracy of

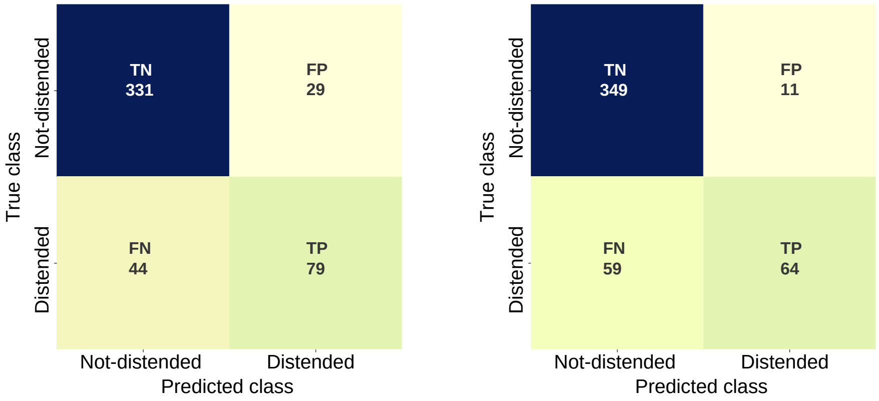
0.73 (95% CI [0.72 – 0.74]). The *Multi-task approach* has a balanced accuracy of 0.78 (95% CI [0.77 – 0.79]) outperforming both the *Classification baseline* and the *Detection approach*. The IoU metric is 0.66 (95% CI [0.65 – 0.66]) for both the *Detection baseline* and the *Detection approach* and decreases to 0.63 (95% CI [0.62 – 0.63]) for the *Multi-task approach*.

These results show that the *Multi-task approach* is the most suitable solution for the considered problem since it has an acceptable level of balanced accuracy and IoU according to prior literature [213, 208]. This conclusion is also supported by taking into account the confidence intervals: the *Multi-task approach* confidence interval range is entirely above the thresholds for both classification and detection, and the balanced accuracy CI does not intersect with the *Detection approach* interval, suggesting that its performances are consistently better [214]. The increase in balanced accuracy value of the *Multi-task approach* is largely influenced by the increase in *sensitivity*. The reason for this increase is likely due to the adjusted classification loss in the *Multi-task approach* introduced to mitigate the unbalanced data problem (see Section 4.2.2). Indeed, considering the confusion matrices in Figure 4.4, we can observe that the *Detection approach* has 59 false negatives (48%), out of a total of 123 images labeled as *Distended*, compared to the 44 false negatives in the *Multi-task approach* (38%). This improvement comes at a cost of a lower *specificity* value that, however, is less relevant than *sensitivity* in the given domain.

4.3.6 Examples

In order to better illustrate how our approaches work, in the following we show some examples of correct and incorrect output.

Figure 4.5 shows two US images that have been correctly classified by both approaches and that are relatively easy to classify by medical experts. Figure 4.5a shows an US image where the femur, the patella and the SQR are clearly visible, and the SQR is thin (*i.e.*, *Non-distended*). On the other hand, Figure 4.5b shows an example of



(a) Multi-task approach

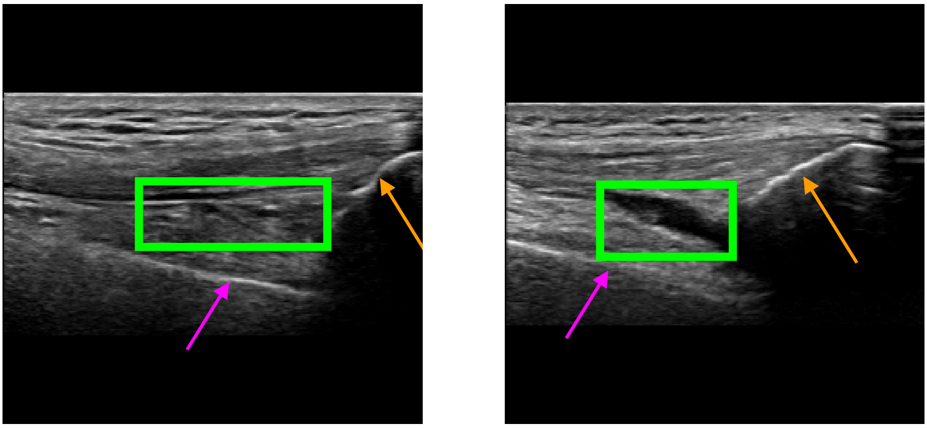
(b) Detection approach

Figure 4.4: Confusion matrices

a *Distended* SQR. In this case, the SQR is clearly thick and hence *Distended*.

Figure 4.6 shows four examples of images that are more challenging to classify even by medical experts. This usually happens when there is noise in the US scan (as in Figure 4.6c) or when the SQR is borderline between *Distended* and *Non-distended* (as in Figure 4.6d). Figure 4.6a is correctly classified by both approaches as *Non-distended*. Figure 4.6b is correctly classified by the *Multi-task approach* but not by the *Detection approach*. Vice versa, Figure 4.6c is correctly classified by the *Detection approach* and not by the *Multi-task approach*. Finally, both solutions wrongly classify Figure 4.6d.

Considering the detection problem, Figure 4.7 shows US images where the two approaches detected the SQR with the lowest and the highest IoU. In Figure 4.7a, the *Multi-task approach* wrongly detects as SQR an image region that is similar to an actual SQR in terms of position and shape, resulting in a very low value of IoU (0.33). In this case, also the *Detection approach* can not reliably detect the right target precisely, and indeed it detects only a small



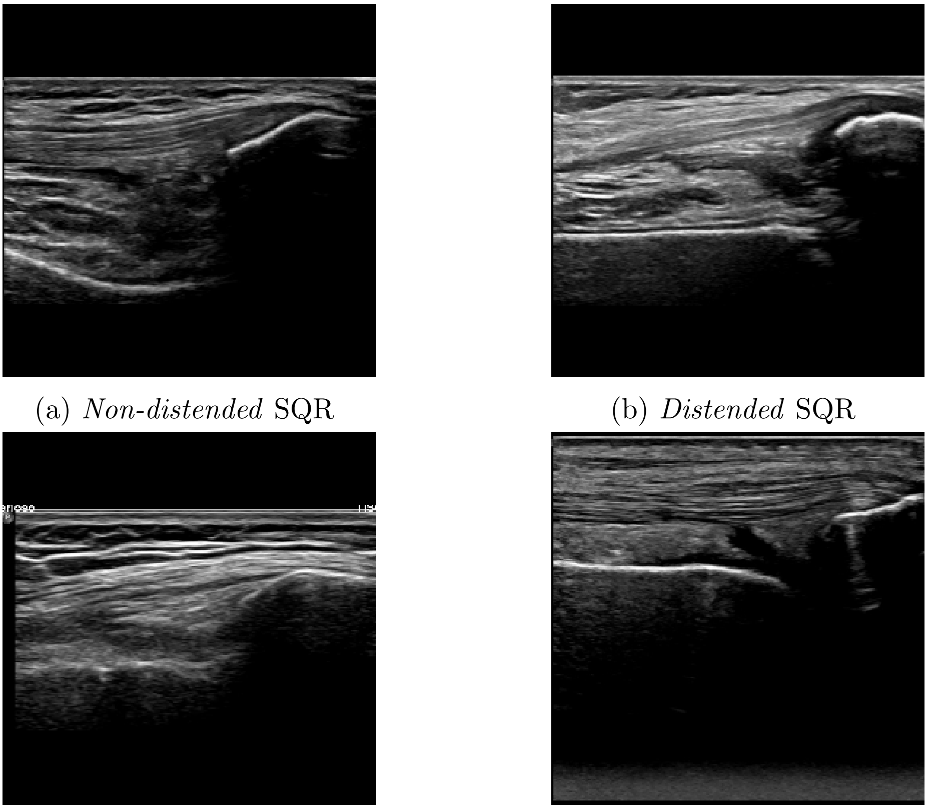
(a) *Non-distended* SQR (b) *Distended* SQR

Figure 4.5: Examples of images correctly classified by both solutions. The purple arrow points to the femur, the orange arrow points to the patella, and the green box indicates the SQR.

portion of the actual SQR (IoU=0.05). Instead, in the example shown in Figure 4.7b the *Multi-task approach* accurately detects the SQR (IoU=0.95), while the *Detection approach* identifies the same area with a lower IoU (0.68).

Figure 4.7c shows the US image for which the *Detection approach* provided the lowest IoU value. The problem is similar to that of Figure 4.7a: a region is erroneously recognized as a SQR because it is similar to a SQR. In this case, the detected bounding box does not overlap with the ground truth, hence the IoU is zero. Instead, the *Multi-Task approach* basically detects the right target (IOU=0.58).

Figure 4.7d shows instead the US image for which the *Detection approach* provided the highest IoU value (0.96). In this case, the *Multi-task approach* identifies the right target less precisely, resulting in an IoU of 0.55.



(a) *Non-distended* SQR

(b) *Distended* SQR

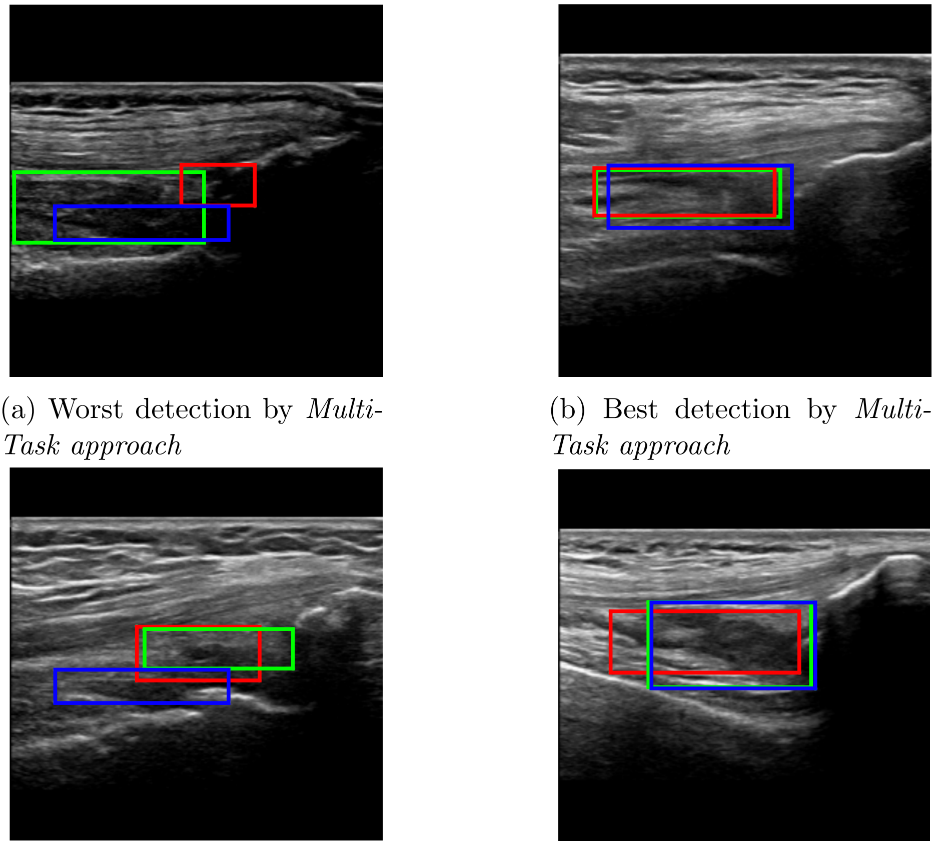
(c) *Non-distended* SQR

(d) *Distended* SQR

Figure 4.6: Examples of images that are intuitively hard to classify.

4.4 Conclusions

In this Chapter we investigate the requirements of a CAD tool that detects joint recess distension from US images can support practitioners in diagnosing hemarthrosis and we frame the problem in terms of a combination of two typical machine learning tasks: classification and detection. Addressing this problem is particularly challenging for a number of reasons, including that the position and the shape of the joint recess may change considerably across different US images,



(a) Worst detection by *Multi-Task approach*

(b) Best detection by *Multi-Task approach*

(c) Worst detection by *Detection approach*

(d) Best detection by *Detection approach*

Figure 4.7: Detection examples. Green represents the ground truth, red and blue the results of the *Multi-Task approach* and *Detection approach*, respectively.

and there can even be borderline cases in which the recess is only partially *Distended*.

We initially proposed a Multi-task learning algorithm which is particularly relevant for two reasons. First, as opposed to baseline solutions based on transfer learning, the balanced accuracy confidence interval of the *Multi-task approach* is completely above the threshold of 0.75 which is reported to be a requirement for a medical test to be "useful" [213]. Hence, the *Multi-task approach* is suitable for our application domain.

Another important property of the *Multi-task approach* is that it yields a substantially higher sensitivity value relative to baselines based on transfer learning. This is particularly important because, in the domain considered, sensitivity should be favored over specificity. Indeed, false negatives (captured by sensitivity) have a worse impact on the patient than false positives (captured by specificity). This is due to the fact that a false positive prediction can lead to raising the practitioner's attention when not needed and, in the worst scenario, can lead to over-treatment (*e.g.*, provide factor VIII when not needed) which generally results in limited negative effects on the patient. Instead, a false negative prediction can lead to under-treatment, which in turn can lead to permanent articular damage [10].

In the proposed solution, the IoU is above 0.5 in more than 82% of the cases. In these cases (and also in many cases in which the IoU is below 0.5) the target SQR is correctly detected, but the detected bounding box is imprecise. There are only a few cases in which the techniques detect the wrong target.

One general limitation of multi-task learning is the difficulty in finding the optimal trade-off among the different tasks. This is particularly relevant in our study, because, due to time and computational power limitations, we were unable to extensively explore the hyperparameter space. This incomplete search limited our ability to identify the best configuration to effectively balance the contributions of each task, resulting in a possible bias towards one of the learned objectives. Another limitation is related to the possibility to use the model prediction to support explainability. Indeed, bounding box, which results from the detection branch and which can be shown to the practitioner, does provide a reliable indication of the model reasoning for the classification.

Although we have achieved some promising results, to effectively implement this system in practice, we need to improve the accuracy of predictions. Currently, there is still a margin of error that could lead to undertreatment in certain cases. As a result, at this stage, the model cannot be relied upon for a comprehensive screening. Instead, it can be used to prioritize cases with more certain indications of joint distension, ensuring that the highest-risk cases are addressed first.

5

Weakly-supervised Anomaly Detection for Ultrasound Images

In the work proposed in Chapter 4, the task of distinguishing between *Distended* and *Non-distended* recesses is addressed with supervised classification. In addition to classification, segmentation is also of utmost importance in medical imaging, as it facilitates the identification of structures or regions of interest, thereby enabling visual guidance for professionals [215]. A major problem of these solutions is the reliance on labeled images, which are scarce, imbalanced between the two classes (*Distended* cases are rarer than *Non-distended* ones) and have a high annotation cost.

In the literature, a common approach to tackle these types of problem is unsupervised anomaly detection [95], in which the model is trained only using normal data samples and is used to identify anomalous samples deviating from the learned distribution. However, as we show in this chapter, these techniques are ineffective in the specific domain considered in our work.

To address the ineffectiveness of unsupervised anomaly detection techniques, we propose a solution inspired by weakly supervised segmentation approaches that have been extensively researched in the segmentation domain, where acquiring the segmentation masks is not always feasible [216]. These approaches rely on *weak* labels that contain partial information compared to the labels used in the supervised approach. In particular, previous work suggests that the use of a *location prior*, in the form of the bounding box of the element of interest, can effectively mitigate the cost of annotation while still providing high accuracy in semantic segmentation [217], referring image segmentation [218], and in medical image segmentation [116, 219]. However, to the best of our knowledge, these approaches have never been applied in the field of anomaly detection.

In this chapter, we present **LoRIS** (Localized Reconstruction-by-Inpainting with a Single mask), a weakly supervised anomaly detection approach that uses the joint recess bounding box as prior knowledge during the inpainting. We also propose *Directional Distance* (DD), a new image similarity deviation metric that yields better anomaly segmentation results than existing metrics, such as Multi-Scale Gradient Magnitude Similarity Deviation (MSGMSD) [220]. Experimental results, conducted on a dataset of 483 images, show that **LoRIS** is more accurate in detecting recess distention when using MSGMSD (image-level AUROC 0.78), outperforming state-of-the-art unsupervised techniques and providing similar results as a previous approach specifically designed for this problem [1]. Instead, considering the segmentation problem, **LoRIS** provides better results when adopting DD (Dice score of 0.35), outperforming existing unsupervised techniques.

5.1 Methodology

After defining the problem (Section 5.1.1), we describe the two main steps of **LoRIS**: *localized reconstruction* (Section 5.1.2) and *anomaly detection* (Section 5.1.3). Finally, Section 5.1.4 describes how to automatically compute the location prior.

5.1.1 Problem Formulation

Hemophilia is a rare disease and its management has improved dramatically in the last decade for two reasons. First, the use of US imaging emerged as a practical solution for the detection of recess joint distention, caused by joint bleeding [14]. A second factor is the increased availability of replacement treatments (coagulation factor VIII and factor IX) and non-replacement treatments [221]. This has led to a reduction in the number of acute bleeding episodes, including intra-articular bleeding [16], which is otherwise a common cause of recess distention. Since in this work, we consider a cohort of patients treated with these drugs, images of *Distended* recesses are rarer than *Non-distended* ones. For this reason, we propose to frame the problem as an anomaly detection task in which a *Distended* recess represents the anomalous case.

Specifically, we address the problem of detecting the distension of the subquadricipital recess (SQR), which is the main recess of the knee joint. Our approach uses images of the longitudinal US scan of the knee joint, which are commonly used to diagnose SQR distention by medical practitioners [14].

5.1.2 Localized reconstruction

The *localized reconstruction module* takes in input an image of the longitudinal US scan of the knee joint and the recess bounding box location prior (see Figure 5.1). The module first inpaints the area in the image defined by the location prior with a black rectangle. Then, it reconstructs the inpainted area using a specifically trained network. One advantage of reconstructing the detected recess area only

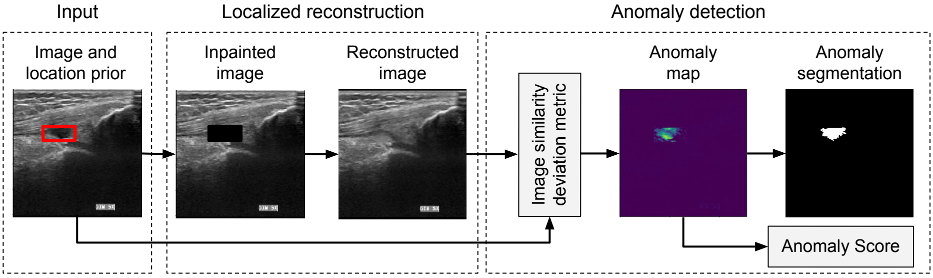


Figure 5.1: **LoRIS** procedure schema

is that this solution avoids reconstructing areas that are of no interest for the given problem and that, due to noise and high inter-patient variability, can be reconstructed imprecisely also for physiological (non-pathological) images.

The network used is a U-Net [222], trained on a single class (images with *Non-distended* recess) to reconstruct the inpainted image while focusing solely on the masked region. This is achieved through skip connections, directly propagating the information from low-level layers to the higher ones, facilitating the reconstruction process by preserving fine details, and maintaining contextual information from the original input. Consistently with previous works [104], we trained the network with the sum of three different losses:

$$\mathrm{L}_{\mathrm{tot}} = \mathrm{L}_{\mathrm{MSGMS}}(I, I_r) + \mathrm{L}_{\mathrm{SSIM}}(I, I_r) + \mathrm{L}_2(I, I_r)$$

where $\mathrm{L}_{\mathrm{MSGMS}}$ is the Multi-Scale Gradient Magnitude Similarity loss, $\mathrm{L}_{\mathrm{SSIM}}$ indicates the structural similarity index loss and the pixel-wise loss L_2 between the original image I and the reconstructed one I_r

At inference time, image reconstruction is achieved in a single iteration that reconstructs the entire masked area. This is in contrast with the approach of using multiple masks, adopted by existing reconstruction-by-inpainting techniques, that iteratively mask and reconstruct portions of the image, finally joining all the reconstructed areas to obtain the entire reconstructed image [104]. The problem with the multiple-masks approach is that, during its iterations, only a portion of the recess could be masked at a time, hence resulting in the image being precisely reconstructed even when the recess is *Distended*. Instead, by using a single mask, the entire recess is inpainted, so it is more likely that it will be reconstructed as *Non-distended* also in *Distended* images, hence revealing the anomaly.

5.1.3 Anomaly detection

At inference time, **LoRIS** runs the localized reconstruction module to obtain the reconstructed image. Then, an *anomaly map* is computed, indicating an anomaly score for each pixel of the original image, using an image similarity deviation metric (see Fig 5.1). An overall anomaly score is computed at image level by average pooling the pixel-wise anomaly scores of the anomaly map. The anomaly is segmented by first selecting the set of pixels in the anomaly map whose value is above a threshold that maximizes the dice score (as in [106]) and then by applying a post-processing step using morphological closing, followed by opening with kernel $3x3$.

In this chapter, we propose a novel image similarity deviation metric called **directional difference** (DD) that is based on the following observation: a *Distended* recess appears in a US image as a thick dark area, whereas a *Non-distended* recess appears as a thin dark line on a lighter background. If an image containing a *Distended* recess is provided in input, we expect the reconstruction to produce an image that resembles a *Non-distended* recess, with the recess bounding box containing lighter pixels, on average, than the original image. The DD metric measures the increase of light intensity for the pixels in the reconstructed image with respect to the original one, ignoring the pixels where the light intensity actually decreases. Formally:

$$\mathrm{DD}(p, p_r) = \max((p_r - p), 0)$$

where p_r is the intensity of a pixel in the reconstructed image and p is the intensity of the corresponding pixel in the original image.

We experimented **LoRIS** also considering alternative image similarity deviation metrics. Some of them are derived from the existing literature on similarity deviation between images, including Gradient Magnitude Similarity Deviation (GMSD), and Multi-Scale Gradient Magnitude Similarity Deviation (MSGMSD) [223]. We also experimented with similarity scoring functions between images by computing their dual, such as the Structural Similarity Index (SSIM). Among all these image similarity deviation metrics, **LoRIS** obtained the best results with MSGMSD in terms of image-wise and pixel-wise accuracy, while best Dice score was obtained using **DD**.

5.1.4 Automatic detection of the recess bounding box

LoRIS requires the recess bounding box as location prior both at training and inference time. The use of (manually annotated) bounding box priors limits the real-world applicability of the proposed approach. To address this issue, we further propose the use of object detection for automatically annotating the bounding box location priors, thus achieving a fully automated pipeline (from image acquisition to anomaly prediction). Note that, also in this case, the object detection has to be trained on *Non-distended* images only to maintain the applicability of the approach in the anomaly detection setting.

5.2 Experimental evaluation

This section describes the experimental methodology (Section 5.2.1), the experimental results in terms of anomaly detection and segmentation performance (Section 5.2.2), and the impact of automatic location prior detection (Section 5.2.3).

5.2.1 Experimental methodology

We used the same dataset used in Chapter 4, containing 483 US images of the knee recess, 123 of which are *Distended*, according to the annotation of a physician who is an expert US reader in this specific field. The same physician also annotated the images with the recess bounding box (the location prior) and the recess segmentation, which is used to compute pixel-wise segmentation accuracy. The images are divided into 5 folds using patient-based splits, thus ensuring that no images of the same patient are simultaneously in the training and test folds. Due to this, the exact number of images in each fold can vary. Approximately, each fold contains 308, 78, and 97 images for the training, validation, and test sets, respectively. Note that the images of *Distended* recesses in the training and validation sets are ignored for the training of the proposed anomaly detection technique. Therefore, for each fold, we use approximately 226, 63, and 97 images in the training, validation, and test sets, respectively.

For what concerns the model training, for each fold, the U-net model was trained for 1000 epochs with an early-stopping criterion of 50 epochs on the validation loss, a learning rate of 0.0001 with Adam optimizer [224], and a batch size of 4. Parameters were selected empirically. We ran the experiments on a Ubuntu Server with a partitioned NVIDIA A100 GPU, 42Gb of RAM, and an AMD EPYC 8-core CPU. The code is publicly available[1].

To assess the accuracy of the anomaly detection, we consider metrics commonly used in the state-of-the-art: Image-level AUROC (I-AUROC) and Pixel-level AUROC (P-AUROC). Additionally, we employ the Dice score as it more accurately evaluates the anomaly segmentation accuracy [225].

5.2.2 Anomaly detection and segmentation results

Table 5.1 compares **LoRIS** with state-of-the-art unsupervised anomaly detection approaches and a previously proposed supervised technique [1]. Considering the unsupervised techniques, recent ones (PatchCore[107], Simplenet [106] and Cflow [108]) yield the best results, with PatchCore having an I-AUROC of 0.701 and a P-AUROC of 0.871. However, unsupervised techniques have a Dice score lower than 0.2, showing that they do not obtain a relevant segmentation of the anomalous region. This is also exemplified in Figure5.2 that shows, for three

[1] `https://github.com/warpcut/LoRIS`

sample images, the segmentation results of various techniques. As shown in the figure, the unsupervised techniques fail in most of the cases to detect the recess area, and, even when they do, they do not approximate the recess accurately. The *multi-task* supervised technique [1] achieves a higher I-AUROC value of 0.780 but it cannot compute the recess segmentation.

Table 5.1: Anomaly detection and segmentation results

Model	**Setting**	**I-AUROC**	**P-AUROC**	**DICE**
RIAD [104]	Unsupervised	0.583 ± 0.083	0.682 ± 0.016	0.051 ± 0.017
InTrans [226]	Unsupervised	0.581 ± 0.053	0.574 ± 0.033	0.028 ± 0.009
Ganomaly [103]	Unsupervised	0.573 ± 0.035	-	-
FAIR [227]	Unsupervised	0.544 ± 0.035	0.668 ± 0.021	0.102 ± 0.012
Cflow [108]	Unsupervised	0.645 ± 0.125	0.864 ± 0.011	0.124 ± 0.049
Draem [105]	Unsupervised	0.547 ± 0.066	0.626 ± 0.041	0.033 ± 0.007
UAE [109]	Unsupervised	0.621 ± 0.068	0.699 ± 0.014	0.061 ± 0.018
Simplenet [106]	Unsupervised	0.68 ± 0.104	0.818 ± 0.01	0.144 ± 0.047
PatchCore [107]	Unsupervised	0.701 ± 0.090	0.871 ± 0.009	0.193 ± 0.066
Multi-task [1]	Supervised	0.780 ± 0.050	-	-
LoRIS+MSGMSD	Weakly-supervised	**0.783 ± 0.050**	**0.932 ± 0.018**	0.263 ± 0.042
LoRIS+DD	Weakly-supervised	0.750 ± 0.100	0.746 ± 0.047	**0.353 ± 0.034**

Table 5.1 also shows the results of two variants of **LoRIS**, when using MSGMSD (**LoRIS+MSGMSD**) and DD (**LoRIS+DD**) as image similarity deviation metrics. The former achieves the best performance in terms of image-level AUROC (0.783) when compared with all other techniques, including *multi-task*. It also outperforms all other unsupervised techniques in terms of pixel-level AUROC (0.932). Taking into account the segmentation ability, **LoRIS+DD** achieves the best results, with a Dice score of 0.353. However, we note that the Dice score is still relatively low, indicating that accurate anomaly segmentation in this domain is particularly challenging. This observation is also supported by the results obtained using UAE [109] which, despite being designed for medical imaging, yields poor results (AUROC of 0.699 and dice of 0.061). Nevertheless, as shown in Figure5.2, while segmentation is not extremely accurate, it approximates the actual recess shape well. More examples of recontructions

and anomaly segmentations can be found in Appendix A.

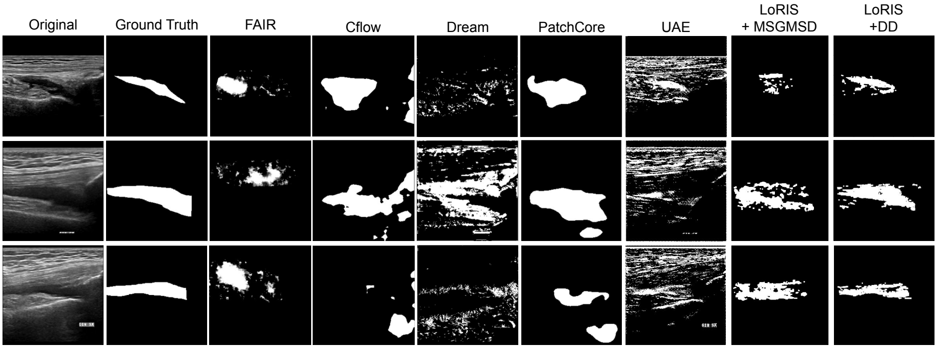


Figure 5.2: Comparison of the anomaly segmentations generated by different techniques

5.2.3 Automated detection of the recess bounding box

For the automated detection of the bounding box location prior, we examine two object detection approaches, Yolo (V5) [201] and CoDETR [228]. We trained the two models on the *Non-distended* images in the training set and measured the performance of **LoRIS+MSGMSD** with the location prior automatically computed by the trained object detection model at test time.

As shown in Table 5.2, YOLO fails to achieve results comparable to the upper baseline represented by the Ground Truth (GT) annotations. Indeed, there is a significant drop in performance: −4.4% in I-AUROC, −5.4% in P-AUROC and −6.2 in Dice score. Instead, using CoDETR, shown to perform better in several domains [228], the results remain comparable with those obtained with GT: −0.7% in I-AUROC, −2.2% in P-AUROC and −2.3% in Dice score.

5.3 Conclusions

The approach proposed in this Chapter is the first anomaly detection technique to use a location prior and to adopt the reconstruction-by-

Table 5.2: Performances of the object detection algorithms and their impact.

	precision	map@50	mAP@75	I-AUROC	P-AUROC	DICE
GT	-	-	-	0.783 ± 0.050	0.932 ± 0.018	0.263 ± 0.042
Yolo-V5	0.954 ± 0.044	0.796 ± 0.052	0.254 ± 0.063	0.773 ± 0.038	0.872 ± 0.033	0.223 ± 0.030
CoDETR	1.0 ± 0.0	0.9 ± 0.035	0.41 ± 0.066	0.776 ± 0.029	0.910 ± 0.038	0.240 ± 0.031

inpainting approach on US images, which are noisy and have high variability. Experimental results show that the technique can separate normal images from anomalous images better than state-of-the-art unsupervised approaches, achieving results comparable to a fully supervised approach proposed in Chapter 4, when **LoRIS+MSGMSD** is used. Instead, **LoRIS+DD** yields the best results for the purpose of anomaly segmentation.

Furthermore, **LoRIS** has two additional benefits with respect to the supervised approach. First, it is trained using non-anomalous data only, and therefore it is more suitable to the target problem domain in which anomalous data is scarce. Second, it provides more anatomically reasonable anomaly segmentations, only requiring the recess bounding box as a location prior. This property will be particularly useful for the continuation of the project and will be discussed in more detail in Section 8.3. We also show that this information can be obtained using a state-of-the-art object detection technique, achieving results comparable to the use of the manually annotated data, and thus achieving a fully automated SQR distension detection pipeline.

One possible limitation of LoRIS is that it would require providing the bounding-box annotations prior to inference. However, by incorporating a detection module, we demonstrated that the weak supervision obtained by the bounding box is no longer necessary once trained. Techniques such as CoDETR [228] can substitute for traditional ground truth annotations, further reducing the work of practitioners. Another limitation is that the quality of the generated recess segmentations results in low overall accuracy as measured by the DICE score. Nevertheless, a qualitative observation of the results indicates that the model's segmentation in the targeted recess areas is often conceptually precise. This intuition is supported by the high value of pixel AUROC. The reason for a low DICE score lies in the recess borders, particularly along anatomical boundaries, which tend to be noisy and difficult to delineate even with manual inspection. This is a known challenge in medical imaging, as subtle and intricate structures often cause difficulty in visual interpretation, leading to inconsistencies in both manual and automated annotations [229].

Finally, the use of the bounding-box supervision could be better integrated in an end-to-end model that is simultaneously trained to detect and inpaint the recess and then perform the reconstruction. Furthermore, the framework was evaluated on RIAD, which is the least performing model among the baselines. An end-to-end model would enable one to directly add the supervision were required by other approaches, and this might lead to significantly better results.

6

Test-time training with contrastive feature reconstruction on ultrasound images

In recent years, test-time training has emerged as an innovative approach in the unsupervised domain adaptation task, with promising applications in various fields. However, its adoption in the medical imaging field remains limited, largely confined to test-time adaptation techniques [230]. Adoption is critical for different reasons: images may have variations in conditions, scanner types, patient demographics, or anatomical structures that affect both the classification and segmentation performance of deep learning models.

Our work leverages the concept of contrastive learning [231] for improving test-time training. The idea of this technique is to learn, from unlabeled data, general-purpose features that are similar in related samples and different in unrelated ones. Previous work has shown the usefulness of contrastive learning in a variety of unsupervised and semi-supervised image tasks [231, 232, 233]. Among others, *ReContrast* [234], which inspired our approach, adopts feature reconstruction contrastive learning in *unsupervised anomaly detection*, demonstrating a good transfer ability to various image domains compared to other unsupervised techniques. However, this recent approach has not been investigated for adapting models at test time.

This paper presents *ReC-TTT* (Contrastive Feature Reconstruction for Test-Time Training), a test-time training approach designed for image classification that leverages techniques from the field of contrastive representation learning in a novel way. The core idea of *ReC-TTT* is to use a pre-trained frozen encoder to generate a discriminative feature representation of the input image. This representation is then used as a positive pair in the learning of the auxiliary task. In particular, during the training phase, two encoders are trained in a supervised manner to classify the images, and, at the same time, a decoder is trained to minimize the differences between the features extracted from the trainable encoders and the ones reconstructed from the frozen pre-trained encoder. The intuition is that during test-time training, the now frozen decoder works as a guide to extract more meaningful information by the trainable encoders.

6.1 Methodology

Our *ReC-TTT* method addresses the problem of domain shift between a given training set, representing the source domain $\mathcal{S} = (X_S, Y_S)$, and a test set from a target domain $\mathcal{T} = (X_T, Y_T)$, where X_S, X_T are spaces containing images and Y_S, Y_T the spaces of corresponding labels. In this setting, we suppose that both domains have the same labels, i.e., $Y_S = Y_T$, but that images have a different conditional distribution, i.e., $p_S(x|y) \neq p_T(x|y)$ where $x \in X$ and $y \in Y$.

Figure 6.1 shows the overall framework of our method. The architecture employed in *ReC-TTT* consists of two trainable encoders $f_{\theta 1}$ and $f_{\theta 2}$, a pre-trained frozen encoder f_{θ_F}, and a decoder g_θ that takes in input the concatenated features extracted from the three encoders. As other TTT approaches, *ReC-TTT* requires two steps. In the first step, our method has access to the source domain and the model is trained to learn a function mapping $X_S \rightarrow Y_S$ using a classification loss ($\mathcal{L}_{CE}$) and an auxiliary loss ($\mathcal{L}_{aux}$). The second step occurs at test time, where our method has only access to the unlabeled target set. In our case, $f_{\theta 1}$ and $f_{\theta 2}$ are updated using only the auxiliary function to learn the new mapping $X_T \rightarrow Y_T$. This partial update enables the model to learn the association in $\mathcal{T}$ while maintaining the knowledge acquired during training on $\mathcal{S}$.

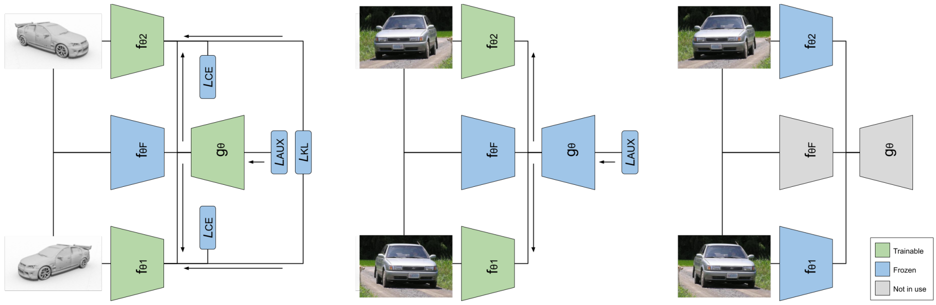
In the following sections, we detail the different components of our method.

6.1.1 Contrastive feature reconstruction

Contrastive learning extracts meaningful representations by maximizing the agreement between the features of different views of the input data during training. In our framework, illustrated in Figure 6.1, this is achieved using two separate encoders. The first one ($f_{\theta 1}$) is updated during training, and hence generates a domain-specific domain representation, while the other ($f_{\theta F}$) is instead frozen and thus generates a domain representation based on a pre-trained network.

The extracted features are then combined into a bottleneck that resembles the last ResNet layer, and subsequently fed into a shared decoder (g_θ) which has the opposite architecture of the encoders. For a fair comparison with previous TTA and TTT works [132, 142], our method uses a ResNet50 backbone for the convolutional feature extractors.

Learning objective. The network is trained using global cosine-similarity [234] between the features at different layers of the encoders and the features at the opposite level of the decoder. Specifically, the model is trained in a cross-reconstruction fashion where the decoder learns to reconstruct the features of the frozen encoder starting with



(a) Train time (b) Test time (c) Inference time

Figure 6.1: **Overview of our *ReC-TTT* framework.** The directional flow of gradients is denoted by the symbol →. $\mathcal{L}_{aux}$ is our cross reconstruction loss, which computes the global similarity between the features of the encoders and the features reconstructed by the decoder, $\mathcal{L}_{CE}$ is the cross-entropy between the predicted classes and the true labels, and $\mathcal{L}_{KL}$ is the Kullback–Leibler divergence between the two predicted distributions. The trainable components of our architecture are depicted in **green**, whereas the frozen components are represented in **blue**. (**a**) illustrates the training phase, where both the encoders and the decoder are trainable. At test-time training (**b**), the decoder is frozen. Finally, (**c**) shows the inference time when the entire network is frozen; modules represented in **gray** are not needed in this phase.

the ones obtained by the trainable encoder, and vice versa, using the following loss:

$$\mathcal{L}_{aux} = \sum_{\ell=1}^{L} 1 - \frac{\langle sg(f_E^\ell), f_D^\ell \rangle}{sg(\|f_E^\ell\|_2)\, \|f_D^\ell\|_2} \tag{6.1}$$

where sg is the stop gradient operation [235] used to avoid propagating the gradient directly into the encoder, f_E^ℓ and f_D^ℓ represent the flattened features of the encoder and decoder respectively at the ℓ^{th} layer, and $\langle \cdot, \cdot \rangle$ is the dot product operation.

In the TTT framework, during training, the network is jointly trained on the two tasks: supervised classification and contrastive feature reconstruction combining the cross-entropy with the auxiliary loss described in Eq. (6.1), as follows:

$$\mathcal{L}_{train} \;=\; \mathcal{L}_{CE} + \mathcal{L}_{aux} \tag{6.2}$$

6.1.2 Encoder ensemble

As shown in Figure 6.1a, our *ReC-TTT* method also leverages an ensemble learning strategy that integrates a secondary trainable encoder ($f_{\theta 2}$) and classification predictor. This encoder takes as input an augmented version of the original image to learn diversified representations of the data and add robustness to the contrastive learning process. The same image is fed to the frozen encoder ($f_{\theta F}$) to generate two contrastive pairs, such that the representations extracted by the decoder should be invariant to the augmentation applied. To avoid introducing information that could artificially facilitate the adaptation to specific domain shifts (as noise), we selected a weak, domain-agnostic augmentation: horizontal flip.

Learning objective. The model is trained with the loss of Eq. (6.2) applied to both encoders. Furthermore, we added a consistency loss between the two predictors, measuring their Kullback–Leibler (KL) divergence, to align the distributions predicted by the two encoders:

$$\mathcal{L}_{train} \;=\; \mathcal{L}_{CE} + \mathcal{L}_{aux} \;+\; \mathcal{L}_{KL} \tag{6.3}$$

Let P and Q be two discrete probability distributions over k classes. The KL divergence is computed as

$$D_{KL}(P \,\|\, Q) \;=\; \sum_{k} p_k \log\left(\frac{p_k}{q_k}\right). \tag{6.4}$$

Adaptation. Algorithm 1 describes how our method is used at test time for adapting the model to data from unseen domains. At this stage, we freeze the shared decoder and, for each test batch, reinitialize the weights of encoders. Afterward, since we have no access to labels for the supervised loss, the layers of the trainable encoders are updated using only Eq. (6.1) for a total of T iterations. For the final inference, the whole model is frozen, and we obtain the final classification by averaging the predictions of two independent encoders. As illustrated in Figure 6.1c to reduce computational complexity during inference, the architecture can be optimized by removing the frozen encoder and decoder, which are no longer necessary for generating predictions.

```
Data: Trained model parameters $\theta_0$, test set $X_T$
Result: Predicted labels $\hat{Y}$
for $param \in \theta_g$ do
    $param.trainable \leftarrow$ False
end
for $batch \in X$ do
    $\theta \leftarrow \theta_0$ ;                          // Initialize weights
    for $iter\ t = 1..T$ do
        Get layers features of batch samples $x$ using model with
          parameters $\theta$;
        $\mathcal{L}_{aux} \leftarrow$ Auxiliary loss between encoders and decoder;
        $\nabla_\theta \mathcal{L} \leftarrow$ Compute gradient of $\mathcal{L}_{aux}$ with respect to $\theta_{t-1}$;
        $\theta_t \leftarrow \theta_{t-1} - \alpha \nabla_\theta \mathcal{L}$ ;        // Update model parameters
    end
    $\hat{y} \leftarrow$ Make prediction using $\theta_T$ on examples $x$;
end
return $\hat{Y}$
```

Algorithm 1: Test-Time Training Algorithm

6.2 Experiments

6.2.1 Experimental setup

Six publicly available datasets were selected for the evaluation. These datasets simulate various types of domain shift: image corruption, natural domain shift, and synthetic to real images.

The Visual Domain Adaptation (VisDA) dataset was designed to pose a new challenge in domain adaptation: from synthetic images to real-world images. This dataset is composed of $152,397$ train images consisting of 2D renderings, $55,388$ validation images extracted from the COCO dataset, and $72,372$ YouTube video frames that compose the test set. All images are labeled into 12 different classes. We evaluated the model's ability to generalize from the training set to the validation set ($\textit{train} \rightarrow \textit{val}$) and from the training set to the test set ($\textit{train} \rightarrow \textit{test}$).

Training protocol. Following previous work, our method employs Resnet50 as the backbone, using 32×32 images for the CIFAR datasets, 64×64 images for the TinyImageNet and 224×224 for the VisDA dataset. The backbone is pre-trained using ImageNet32 [236] for the first one and ImageNet [80] for the latter. Following existing literature, all CIFAR models were trained for 300 epochs with SGD as optimizer, a batch size of 128, and an initial learning rate of 0.1 with a multi-step scheduler, decreasing the learning rate by a factor of 0.1 every 25 epochs. In contrast, for VisDA, the model was trained for 100 epochs, a batch size of 64, and a learning rate of 0.001 without a scheduler.

Inference. At test time, the shared decoder is frozen, while the rest of the network is trained with SGD and a learning rate of 0.005 for CIFAR datasets and 0.0001 for VisDA, using only the auxiliary loss. Similarly to previous approaches, we reset the weights after each batch, hence enabling the consequential processing of batches with different domain shifts.

The experiments were run on a Ubuntu server with an `NVIDIA A100` GPU, 42Gb of RAM, and an `AMD EPYC` 8-core CPU. The code is implemented in python3 with `PyTorch` 1.12.0.

6.2.2 Empirical results

Comparison with state of the art

Our model was compared with seven recent approaches: ResNet50 [178] as the baseline, trained with the same strategy as our method, but only on the supervised task, PTBN [135], TENT [136], TTT++ [142][1], TIPI [137], ClusT3 [132] and *NC-TTT*[143]. For a fair comparison, all TTA methods were evaluated on the same pre-trained ResNet50, while TTT approaches were trained using the same ResNet50 base architecture and the same training strategy.

CIFAR-10 corruptions. Table 6.1 shows the comparison on the CIFAR-10C dataset of the different state-of-the-art methods, the baseline, and our approach. It is noticeable that *ReC-TTT* outperforms on average all previous methods, with a gain of 1.46% on TTT++ and 36.2% on the baseline. Also, our method is the only one able to outperform the baseline for the natural domain shift (CIFAR 10.1, see last line in Table 6.1). As discussed in previous papers [132], other techniques perform worse than the baseline possibly due to the small domain shift between CIFAR-10 and CIFAR-10.1. Instead, *ReC-TTT* can capture this small domain shift thanks to more robust training, thus achieving better performances, 5.5% higher than ClusT3 and around 3% better than *NC-TTT*. To be consistent with the other experiments, we report the performance with 20 adaptation iterations obtaining a gain of 0.27% of AUROC score, while without adaptation our model achieves an even better AUROC of 90.18 (see Section 6.2.2). A common limitation of TTT methods is that they are subject to high variability. To investigate this aspect, we repeated the experiments three times with different random seeds (Table 6.1 reports the average among three runs). The results show that the performance of TTT++ can vary by ±5.05, and those of ClusT3 by ±2.62, whereas *ReC-TTT* yields more consistent results with smaller variations (i.e., ±1.18).

[1] Due to reproducibility issue, TTT++ results were extracted from the latest reported results [132].

CIFAR-100 and TinyImageNet-C corruptions. For the sake of the generability, the hyper-parameters used in the training of *ReC-TTT* for CIFAR-100C and TinyImageNet-C are the same as those used for CIFAR-10C with the only difference that, for CIFAR-100C, the best results are obtained when all the layers are trainable. Figure 6.2 shows that *ReC-TTT* again outperforms the other techniques on both datasets, with a gain of 29.47% when compared to the baseline for CIFAR-100C and a gain of 14.46% for TinyImageNet-C. This demonstrates the robustness of our method to adaptation settings involving a large number of classes.

VisDA. When training on VisDA, *ReC-TTT* achieves the best performance when all the layers of the encoders are trainable, and with 20 iterations of adaptation. Figure 6.2 also reports the results for *train* → *val* and *train* → *test* for VisDA. *ReC-TTT* performs better than all other approaches in *train* → *val* except *NC-TTT* while TIPI and *NC-TTT* show the best results for *train* → *test*. It is worth mentioning that *train* → *val* and *train* → *test* model the same synthetic-to-real domain shift, but using two sources of real images with different characteristics. For instance, the ratio of images for each class varies greatly, and images from the test set are obtained from video frame crops and may thus be blurry, etc.

Visualization of the adaptation

To better understand the effect of adaptation, we consider Figure 6.3 showing the t-SNE plots of the test features before adaptation and after different numbers of iterations for two corruptions types: Brightness and Contrast. In the top row (Brightness), we can observe that *ReC-TTT* obtains a good separation of the features for the different classes also without iterations (AUROC of 92.31). Successive iterations further separate the clusters but have a marginal impact on performance (AUROC of 94.03 at iteration 20). The results are different in the bottom row (Contrast). In this case, without adaptation, most features overlap, without a clear separation, and, indeed, *ReC-TTT* reaches an AUROC of 48.14. With successive iterations,

Corruption Type	**ResNet50**	**PTBN**	**TENT** (ICLR2020)	**TTT++** (NEURIPS2021)	**TIPI** (CVPR2023)	**ClusT3** (ICCV2023)	***NC-TTT*** (CVPR2024)	***ReC-TTT***
Gaussian Noise	23.65	57.49	57.67	75.87 ±5.05	71.90	**75.81** ±2.62	75.24 ±0.12	71.97 ±1.18
Shot Noise	27.68	61.07	60.82	77.18 ±1.36	78.24	77.32 ±2.14	**77.84** ±0.15	75.44 ±1.02
Impulse Noise	32.00	54.92	54.95	**70.47** ±2.18	59.64	67.97 ±2.78	68.77 ±0.15	69.28 ±0.27
Defocus Blur	38.73	82.23	81.39	86.02 ±1.35	84.67	88.10 ±0.20	88.22 ±0.04	**89.56 ±0.18**
Glass Blur	36.49	53.91	53.45	69.98 ±1.62	67.62	60.47 ±1.72	**70.19** ±0.18	69.38 ±0.73
Motion Blur	49.85	78.38	78.13	85.93 ±0.24	82.39	84.99 ±0.49	86.82 ±0.10	**88.94 ±0.03**
Zoom Blur	44.58	80.87	80.56	88.88 ±0.95	85.01	86.76 ±0.29	88.36 ±0.10	**89.65 ±0.27**
Snow	65.39	72.06	71.46	82.24 ±1.69	80.68	81.46 ±0.39	84.42 ±0.07	**86.75 ±0.44**
Frost	48.55	68.68	68.81	82.74 ±1.63	82.12	80.73 ±1.25	84.80 ±0.06	**86.83 ±0.59**
Fog	58.81	76.32	75.94	84.16 ±0.28	76.05	82.52 ±0.25	86.81 ±0.12	**88.87 ±0.33**
Brightness	84.72	85.38	84.87	89.97 ±1.20	88.96	91.52 ±0.24	92.52 ±0.04	**94.03 ±0.24**
Contrast	25.38	81.27	80.65	86.60 ±1.39	76.49	82.59 ±0.92	87.84 ±0.11	**89.56 ±0.48**
Elastic Transform	60.90	67.76	67.21	78.46 ±1.83	77.25	80.04 ±0.35	80.23 ±0.06	**81.66 ±0.32**
Pixelate	39.25	69.59	69.22	82.53 ±2.01	82.67	81.69 ±0.58	81.93 ±0.22	**82.13 ±0.34**
JPEG Compression	64.96	66.50	66.17	81.76 ±1.58	79.39	**81.58** ±1.18	78.49 ±0.09	79.69 ±0.12
Average	46.73	70.43	69.93	81.46	78.21	80.67	82.17	**82.92**
CIFAR 10.1	89.00	86.40	85.30	88.03	85.70	83.77	86.40	**89.27**

Table 6.1: Performance comparison with state-of-the-art on CIFAR-10C and CIFAR10.1 (%).

the cluster separation improves (AUROC of 89.56 at iteration 20) thus demonstrating the effectiveness of our adaptation technique on the extracted features. On the other hand, for a limited number of samples that are wrongly classified before the adaptation, the distance of these samples to the true class increases with the number of iterations.

Robustnenss to smaller batch sizes

As demonstrated in [137], most domain adaptation approaches suffer from the need for large batch sizes to achieve competitive results. Most methods are usually evaluated with batches that have a size of 128 or more. This is a limitation in the application in which it is not possible to collect large batches before computing the inference. For this reason, we compared the performances using different batch sizes (8, 32, 64, 128) for the CIFAR-10C dataset. Table 6.2 reports the results of this study showing that most of the SOTA approaches lose up to 7% of AUROC when the number of samples is lower than the number of available classes (a performance degradation is also reported for TIPI in [137]). In contrast, the performance loss of *ReC-TTT* is less than 2% even with the smallest batch size.

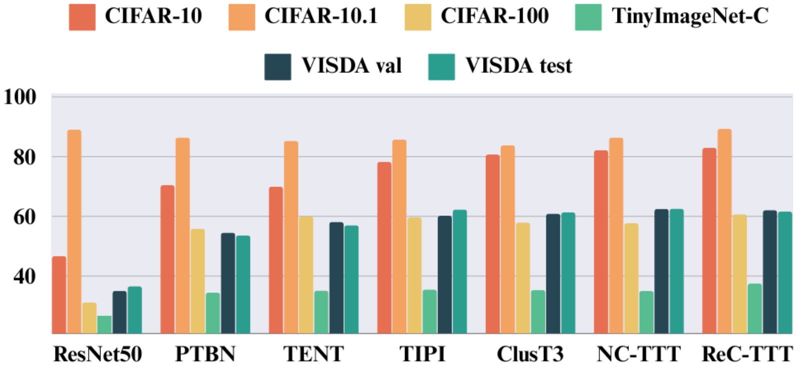
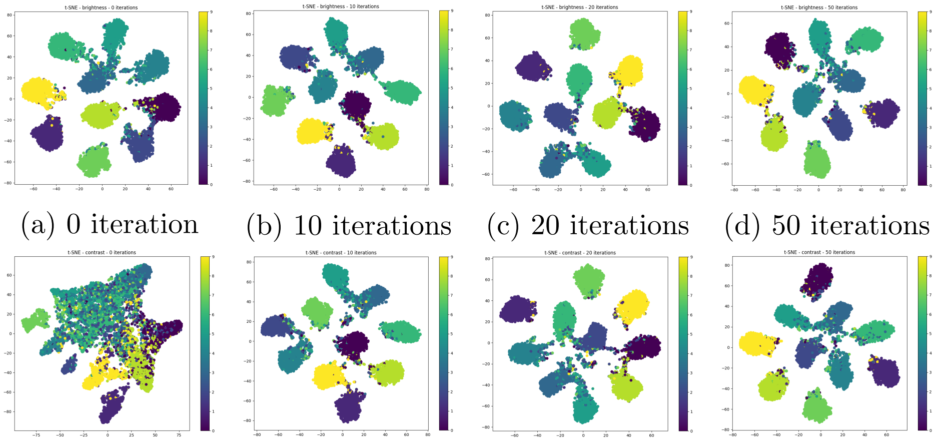


Figure 6.2: Quantitative results, compared to the state-of-the-art, on the CIFAR TinyImageNet-C and VisDA datasets (%). A detailed report for CIFAR-100, TinyImageNet and VisDA is provided in Appendix B

.

Which layers to adapt?

Previous studies suggest that the selection of layers that are updated by the auxiliary task at test time can affect performance [142, 237, 132]. Table 6.3 reports the results of the adaptation of different layers on CIFAR-10C, having the best performance when updating only the first three ResNet blocks. While the difference in performance between three or four trainable layers is negligible (+0.17%), adapting the first two layers yields a reduction in performance of 2.3%, and using the first layer only results in a drop of performance of 18.13%. Differently, for CIFAR-100C and VisDA, we obtained the best results by updating all four layers of the trainable encoders. This can be explained by the greater challenge posed by these datasets, i.e., the larger number of classes for CIFAR-100C and the harder synth-to-real domain shift for VisDA, requiring adaptation of features in deeper layers.



(a) 0 iteration (b) 10 iterations (c) 20 iterations (d) 50 iterations

(e) 0 iteration (f) 10 iterations (g) 20 iterations (h) 50 iterations

Figure 6.3: t-SNE plot of features after different adaptation iterations (0, 10, 20, 50) for the *Brightness* (top row) and *Contrast* (bottom row) corruptions of CIFAR-10C. The adaptation at test time helps separate the features of examples from the same class (represented by color).

Number of adaptation iterations

Another important aspect of test-time adaptation is the number of iterations needed at test time to obtain the best results. In line with previous studies [139, 132] Figure 6.4 shows, for the corruption types of CIFAR-10C, that the best results are obtained after 20 iterations. Successive iterations do not yield better results, on average. The same image for CIFAR-100 corruptions can be found in supplementary material. The same finding emerges from the other datasets with the only exception of CIFAR-10.1 where, as per previous experiments [139, 132, 143], adaptation tends to degrade the performances (90.18% of AUROC without adaptation, 89.27% of AUROC after 20 iterations).

Batch size	PTBN	TENT	ClusT3	*NC-TTT*	*ReC-TTT*
8	61.97	62.11	73.73	79.75	**81.68**
32	68.46	68.46	80.14	81.80	**82.54**
64	69.45	69.54	80.38	82.01	**82.84**
128	70.43	70.09	80.67	82.43	**82.92**

Table 6.2: **Robustness to the batch size.** Qualitative performance of different approaches on CIFAR-10C (%) for different bacth sizes used during training.

Trainable layers	Impulse Noise	Brightness	Pixelate	Average
1 layer	16.12	93.26	34.83	64.79
2 layers	61.86	94.00	76.46	80.62
3 layers	69.25	94.06	82.03	**82.92**
4 layers	69.56	93.78	81.57	82.75

Table 6.3: **On the impact of the training different layers.** Performance comparison of training different layers of our approach on CIFAR-10C (%).

On the contrastive loss performances

To show the impact of our contrastive approach we implemented a TTT method based on the *SimSiam* [235] framework. This solution only compares the features at the bottleneck level and is based on a single encoder, followed by a projection head and a predictor. The model was trained with the Cross-Entropy loss and the *SimSiam* loss as auxiliary task. As reported in the paper presenting the *SimSiam* technique [235], the loss is computed as the negative cosine similarity between *i*) the features of the projector (f_E) extracted by the original image and *ii*) the features of the predictor (f_P) of the augmented version of the image with a stop gradient on the predictor features. To have a fair comparison with *ReC-TTT*, we also used horizontal

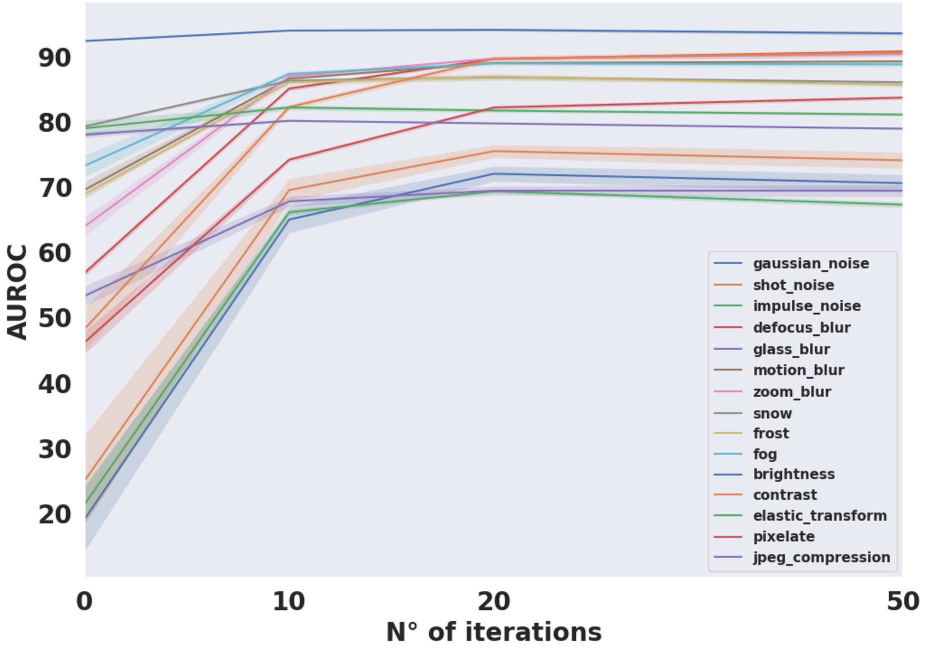


Figure 6.4: **How many iterations are needed for adaptation?** Performance (AUROC) obtained by our method with different number of adaptation iterations on CIFAR-10C. For most corruption types, our method provides a significant boost within few iterations and remains stable when the number of iterations is increased.

flip as augmentation. During the adaptation phase, we adopted the same auxiliary loss to adapt the encoder features for a total of 20 iterations.

Table 6.4 shows that the *SimSiam* contrastive learning approach, although achieving some good adaptation performances, does not achieve the same results as *ReC-TTT*. A possible reason for this result is that *SimSiam* cannot fully capture the domain shift, which is hidden in the whole representation and not only at the bottleneck level. This is the main difference with *ReC-TTT* that instead compares features at different layers.

	Impulse Noise	Brightness	Pixelate	Average
SimSiam	56.40	82.92	68.69	69.77
ReC-TTT	69.28	94.03	82.13	**82.82**

Table 6.4: **On the contrastive loss.** Qualitative results using SimSiam contrastive approach on CIFAR-10C (%).

	Impulse Noise	Brightness	Pixelate	Average
One encoder	65.19	93.17	80.36	81.07
Two encoders	69.28	94.03	82.13	**82.82**
One encoder (Inference)	67.54	93.31	80.89	81.59

Table 6.5: **Using one *vs.* two encoders.** Qualitative results on different configurations of our approach, on CIFAR-10C (%).

Impact of removing the second trainable encoder

We evaluated the effectiveness of *ReC-TTT*'s ensemble learning strategy, which employs two trainable encoders, by comparing it with the base architecture with a single encoder. Table 6.5 shows the performance for three different corruptions and the average among all 15 corruptions available in CIFAR-10C. We observe that using two encoders performs better than having a single one in all cases. This demonstrates the effectiveness of our ensemble learning approach to stabilize training and provide a more robust prediction. To address the increased computational complexity introduced by our model, we investigated the impact of using only one encoder during inference when constrained by performance requirements. While this approach results in a slight reduction in performance, it still yields better results when compared to training the model without the ensemble architecture, remaining competitive with the other SOTA approaches.

6.2.3 Preliminary results on US images

As a preliminary evaluation on US images, we tested our approach on the knee-elbow dataset described in Section 2.7.2. In this scenario, we compared *ReC-TTT* with a set of TTA methods based on the same ResNet50 architecture: ResNet50, PTBN [135], T3A and T3A+BN [238] TENT [136] and TIPI [137]. To train the models, we used the knee SQR dataset, while the target domain is the elbow OLR dataset.

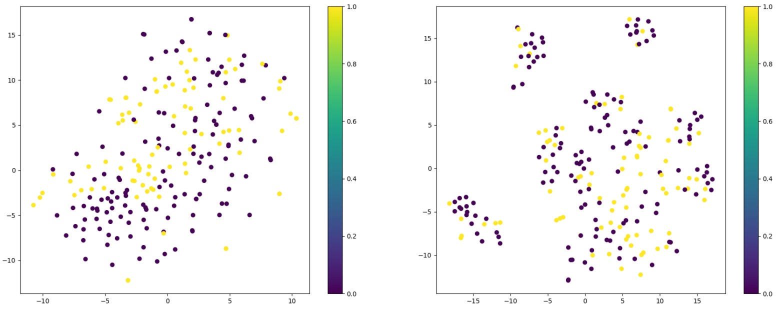
The results obtained are reported in Table 6.6, these results show the balanced accuracy obtained by different methods of TTA and TTT. We can see how the model trained on knee images and directly applied on the elbows fails to classify the images, reaching only a balanced accuracy of 61.00, while all adaptation approaches provide significant improvement, *ReC-TTT* has the highest performance, obtaining a balanced accuracy of 70.49. These preliminary results do not allow for the implementation of the tool for use and still need to be correctly tuned. However, it shows the potential of these approaches in such a scenario. Figure 6.5a shows that before adaptation features are not well separated. After the test-time training step (Figure 6.5b), there is a minor improvement in the separation, but, as confirmed by the numerical results, it is not sufficient to classify the images with high accuracy.

	ResNet50	**PTBN**	**TENT**	**T3A**	**T3A+BN**	**TIPI**	***ReC-TTT***
SQR-OLR	61.25	62.31	63.74	62.96	64.30	63.09	**70.49**

Table 6.6: Performance comparison with state-of-the-art on knee to elbow dataset (%).

6.3 Conclusions

This Chapter addressed the problem of domain shift between training and test data under the Test-Time Training framework. We presented *ReC-TTT*, a novel TTT approach based on contrastive



(a) 0 iterations

(b) 10 iterations

Figure 6.5: t-SNE plot of features before and after adaptation on the knee-to-elbow dataset

feature reconstruction that can efficiently and effectively adapt the model to new unseen domains at test time. Through a series of extensive experiments, we demonstrated that our model outperforms other state-of-the-art approaches across diverse datasets subject to different distribution shifts. An important limitation of previous approaches is their need for large batches of test samples to correctly adapt the model. Our results show that *ReC-TTT* is more robust to this factor, even when the number of classes is greater than the number of available samples. Furthermore, we highlight the robustness of our method against training variability, typically observed in current TTT approaches. Another key advantage of our approach is that it requires tuning few hyper-parameters at test time, specifically, the layers to adapt, the learning rate, and the number of adaptation iterations. Finally, our method was evaluated on knee and elbow musculoskeletal US images, demonstrating promising results; however, these results are not yet sufficient for practical implementation in the CADET system. These results underscore the potential of our method to improve the applicability of TTT in various domains under challenging conditions. As demonstrated in the literature, this allows, especially in the medical field, to solve several problems caused by privacy and security restrictions, which limit the sharing of data between different centers that use different machines. Furthermore, for similar problems, such as in the case of the adaptation to elbow images, it is possible to save time on the images to be annotated, obtaining working models in advance and saving the time of doctors.

We must highlight that our approach, like all TTT models, requires training the entire model and the auxiliary task on the source dataset. This results in a longer training procedure compared to TTA approaches and limits the applicability to the availability of the source data. Secondly, the proposed architecture introduces multiple encoders that might affect the computational and memory requirements during training and adaptation. This is partially solved during inference: these additional components are not required. Lastly, although we validated our method on three classic benchmarks and on the knee-to-elbow dataset, further evaluation should be performed on different pairs of datasets, for example from US images collected with the hospital probes to those collected with the portable US probes. This would enable better performance on the data actually collected by the patients with the portable probes, without requiring the whole annotation process to train a device-specific model.

7

PRACTICE: an intelligent healthcare platform

The application of Artificial Intelligence (AI) methods in the medical domain is a research area investigated by a large number of research groups, due to its potential to revolutionize diagnosis, treatment, and patient care [239].

Combining these advances with the latest trends in the Internet of Things (IoT) makes it possible to build advanced remote monitoring systems taking advantage of sensing devices such as wearable devices, physiological sensors, and smart home sensors [240]. Such systems have the goal of continuously and unobtrusively monitoring the health status of a patient with the long-term objective of improving the patient's quality of life and reducing health system costs.

Most of the existing studies in this area mainly focus on the data analysis aspects that are indeed crucial to provide clinicians with correct and complete information about the patient's health status. However, real-life deployment of these telemedicine systems requires the development of several tools that are rarely investigated in research papers. For instance, several medical domains require the monitored patient to collaborate in data collection (*e.g.*, self-collecting data) and this requires user-friendly applications. Similarly, clinicians who receive AI-processed data from their patients require user-friendly applications that help them analyze the results to make informed decisions. Furthermore, in supervised settings, clinicians also need accurate and easy-to-use annotation tools that can be quickly adapted to research needs.

In this chapter, we describe PRACTICE (Pilot on Remote AutomatiC ulTrasound scan analysIs for hemophiliC patiEnts), a distributed healthcare system designed in collaboration between computer scientists and clinicians to support the application of AI methods in the hemophilia domain. In PRACTICE, each hemophilic patient is provided with a portable ultrasound system. When necessary (*e.g.*, a routine check or in case of pain), the patient uses the probe to acquire US images of the joints that are automatically transmitted to the specialized center where a medical practitioner remotely assesses the presence of joint bleeding supported by state-of-the-art AI methods (such as [1, 241]).

7.1 Requirements

PRACTICE is the result of a multi-year collaboration between two teams of researchers, one from the Computer Science Department of the University of Milan, and the other from the Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, also affiliated with the Department of Pathophysiology and Transplantation of the University of Milan. The collaboration involved multiple funded projects and various research goals, with the overarching objective of supporting the diagnosis process and the follow-up monitoring of joint recess effusions in patients with hemophilia. In this context, the key functional requirements of the platform are:

- Supporting medical practitioners in the diagnosis process and the follow-up monitoring of joint healthcare of the patients through an interactive computer-aided diagnosis tool that shows ultrasound images collected by the patients and estimates the presence of joint recess effusion.

- Providing guidance to patients with hemophilia for the self-acquisition of US scans of their joints using a portable ultrasound probe through an application running on a tablet device that detects anatomical markers of the joint and interactively instructs the user on how to move the probe to correctly scan the recess.

- Facilitate the practitioners in annotating the presence of recess effusion and outlining the recess in the images collected by the patients in order to train the computer-aided diagnosis tool to better recognize recess effusion.

In addition to the medical practitioner and the patient, we also identify two supporting figures, along with their roles:

- The system administrator that manages the users of the platform, assigns the annotation tasks to the medical practitioners, and monitors the completion progress of the annotation tasks.

- The data scientist who uses the annotated images to train the machine learning models.

There are also three non-functional requirements that are relevant for the system design:

- The entire decision process, starting with the acquisition of the US images by the patient and concluding with the determination of the diagnosis by the medical practitioner using the computer-aided diagnosis tool, should not have a longer duration, for the physician, than the usual practice, with the patient going to the hospital for an in-person visit.

- Since the process involves the remote acquisition of US images, their transmission to the hospital servers, and their usage in the annotation system, the training of the machine learning model, and the computer-aided diagnosis tool, it is crucial to guarantee the patients' privacy at all stages of the process.

- Given the research-oriented nature of the project, the data scientist can be interested in exploring various ML models. This requires high flexibility in the data annotation process.

7.2 System architecture

Figure 7.1 shows PRACTICE system architecture. The system is composed of the PRACTICE server, the hospital ultrasound device, the GAJA app running on Windows tablet computers and connected to a portable ultrasound probe, and two web applications: CADET and ATOM. In this Chapter, we will focus on GAJA, to which I made a significant contribution. The CADET and ATOM components, where my involvement was less substantial, are described in Appendix C.

The hospital ultrasound device is a closed system that does not have a publicly available Software Development Kit (SDK). This means that it is not possible to develop ad-hoc applications using the hardware of the ultrasound device. To the best of our knowledge, this is common for most ultrasound devices. Therefore, we integrated this device by leveraging its pre-installed application and configuring it so that, at the end of each visit, it automatically saves the media (images and videos) in a folder on the PRACTICE server. A daemon running on the PRACTICE server watches for changes in that folder and, when it observes a new file, loads the media and its associated metadata (*e.g.*, date of visit) on the database (*main-DB*) through *main-API*, a set of REST APIs implemented through a Node server.

The other three clients (GAJA, CADET, and ATOM) interact directly with *main-API* to store and retrieve data from *main-DB*. All three clients also share a common problem: preserving patients' privacy. To address this issue, the PRACTICE system adopts a pseudonymization approach: all data and media related to a patient are associated with a pseudo-identifier as soon as they are stored in the *main-DB*. All operations related to pseudonymization are implemented by the *pseudonymization-API*, a set of rest APIs that store data in the *identities-DB*, a separate database with higher security (restricted access). In the following processing, the media is associated with the pseudo-identifier, unless the real patient's name is required (*e.g.*, by the practitioner during a visit). In these cases, client applications can access the name through *pseudonymization-API* that implements a role-based access control policy (*e.g.*, the practitioners can access the patients' names, while data scientists cannot).

Finally, there are two other components worth mentioning. The first is *ML-API*, which provides access to the machine learning models through a set of REST APIs available only for local calls and implemented in Python. The second is a set of instances of various annotation tool services. As detailed in Section C.2, ATOM orchestrates various third-party annotation tools, each running with its own instance (and possibly its own database) and interacting with *main-API*.

7.3 GAJA: Guided self-Acquisition of Joint ultrAsound images

One limitation of the current approach is that it is not always possible for the patient to attend frequent visits (*e.g.*, due to the distance from the specialized center). Similarly, frequent and urgent visits can be hard to manage by the specialized center for a number of reasons, including the limited availability of medical personnel and the costs. In order to address these issues, the University of Milan and the Policlinico of Milan are designing a telemedicine system for at-home joint bleeding diagnosis. The idea of the system is that each hemophilic patient is provided with a portable ultrasound probe

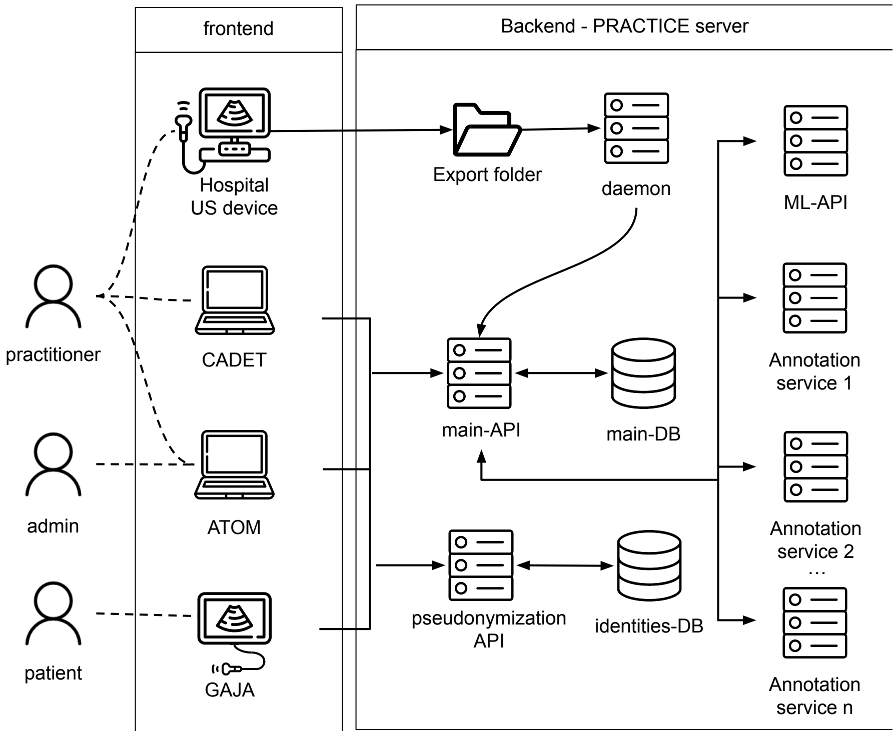


Figure 7.1: Overview of PRACTICE architecture

connected to a portable computer. When necessary, as a routine check or in case of pain, the patient uses the probe to acquire US images of the joints[1] and sends them through the computer to the specialized center where a medical practitioner remotely assesses the presence of joint bleeding, supported by a CAD tool using techniques already proposed in the literature for this problem [1, 241].

One of the main challenges of the system is that the acquisition of ultrasound images is operator dependent, so it is unclear to what extent the patients can acquire images that are suitable for remote diagnosis. This problem has been addressed in the literature with two

[1] We use the term "patient" to denote the person in charge of acquiring the US images but actually it can be the patient or a caregiver.

different approaches: to teach the patients how to acquire the image so that they can repeat the process without additional support [34] or to guide the acquisition in real-time with remote support by a medical practitioner [37]. The limit of the former approach is that patients tend to forget how to acquire the images [41], while the latter approach is time-consuming for the medical practitioners.

To overcome the limitations of existing approaches, we present GAJA (Guided self-Acquisition of Joint ultrAsound images), an application that provides an automated guiding system to support the patient in the acquisition of joint US images. GAJA is designed to combine the benefits of existing solutions: on one side, it guides the patient in real-time during the acquisition process, on the other side it does not require the practitioner to remotely supervise the acquisition process. Currently, GAJA is a working prototype that supports the acquisition of knee joint US images.

7.3.1 Interaction design

GAJA was designed by a multi-disciplinary research team involving computer scientists and medical practitioners. The *Automate-Guide-Remind* design principle was defined in the process and a collaborative interaction approach was adopted.

"Automate-Guide-Remind" design principle

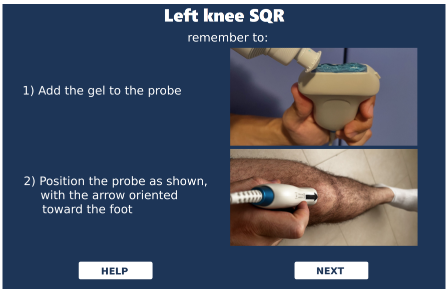
Previous papers show that learning to self-acquire US images is difficult and that patients tend to forget how to use and position the probe after some time. We conjecture that this is partially due to the large number of actions that the patient is required to complete and that affect the successful acquisition of US images: the probe positioning on the body, its inclination, the joint flexion, putting the gel on the probe, and setting the probe parameters. To mitigate this problem we introduce the *Automate-Guide-Remind* design principle. According to this principle, as many actions as possible should be **automated** so that the patient is not in charge of them. The actions that cannot be automated, and that require extensive practice and deep domain knowledge, which usually only medical practitioners acquire during their training, should be **guided**, meaning that the system should provide automatic instructions in real-time on how to use the system. For the remaining actions, which cannot be automated or guided, clear **reminders** should be automatically provided by the system. These actions should only include those that are easy to explain to the patient and that the patient can easily complete.

To implement the *Automate-Guide-Remind* design principle we identified the set of all actions required to successfully acquire a US image, and divided them into three classes:

- **Automated actions**. This class contains actions that can be totally automated and hence are not in charge of the patient. For example, in GAJA the probe parameters (scan depth and gain) are tuned by the practitioner in a setup phase and saved as presets for each scan. During self-acquisition, these parameters are automatically loaded without intervention by the patient.

- **Guided actions**. These are the actions that the patient does while guided in real-time by the system. In GAJA they include the fine positioning of the probe on the body as well as the exact joint flexion.

- **Reminded actions**. These actions are performed independently by the patient, possibly after initial training and with reminders provided during use. In GAJA these actions include general probe usage (*e.g.*, apply the gel on the probe) as well as scan-specific coarse positioning. For example, in the subquadricipital recess (SQR) longitudinal scan the probe should be centered on the leg and parallel to the femur. The patient learns to use the probe and to position it during an initial physical visit with the medical practitioner (the setup step, see Section 7.3.1). When patients need to use GAJA independently, they can access quick reminders as well as a detailed video explanation (by clicking the 'HELP' button) (*e.g.*, see Figure 7.2).



Left knee SQR

remember to:

1) Add the gel to the probe

2) Position the probe as shown, with the arrow oriented toward the foot

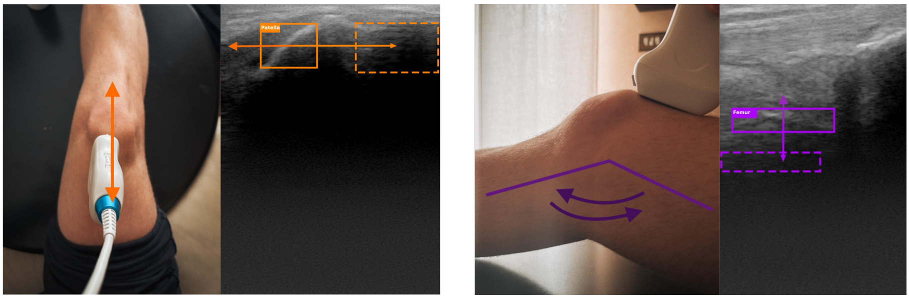
Figure 7.2: Instructions provided before the self-acquisition

Based on this design criterion, in the specific case of the SQR longitudinal scan, we identified two **Guided actions** shown in Figure 7.3 where the solid orange boxes represent the current patella position, while the dashed boxes represent the target area where the patella should be positioned. Similarly, the purple boxes represent the femur. The two **Guided actions** are: positioning the probe with the correct distance from the knee (see Figure 7.3a) and flexing the knee to the right angle (see Figure 7.3b). These actions are particularly important because even small errors can make the acquired US image unsuitable for the diagnosis. In particular, as the probe gets closer to the knee (see Figure 7.3a left), the patella (rigid orange box in Figure 7.3a right) moves right in the scan. Similarly, increasing the knee flexion angle (see Figure 7.3b left) moves the femur down in the scan. We empirically selected these two actions based on the experience of the medical practitioners in our research team.

Collaborative interaction approach

GAJA adopts a collaborative interaction approach between the medical practitioner and the patient. The approach consists of a *setup* step in which the medical practitioner and the patient collaborate in person and a *self-acquisition* step in which the patient independently acquires the image.

Setup step: The setup step is conducted during an in-person clinical visit by an expert medical practitioner who trains the pa-



(a) Probe distance from the knee

(b) Knee flexion

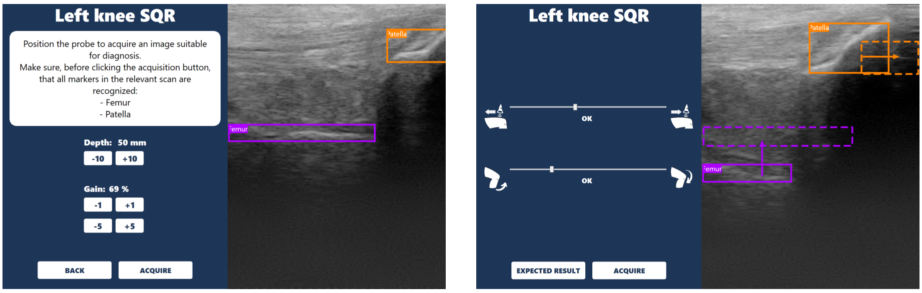
Figure 7.3: Effect of moving the probe or the knee angle

tient (preliminary results show that a training session of about 10 minutes is sufficient) and collects a **reference image** for each target joint scan[2]. The collection of reference images is particularly relevant because the correct probe positioning may vary between patients having different physical characteristics and health conditions, hence it is important to personalize the probe position for each patient.

During the training, the practitioner shows how to use the system and provides basic instructions on how to coarsely position the probe and how to follow the guidance instructions. During the reference image collection, scan-specific anatomical markers are automatically extracted from the US image using object detection techniques. For example, Figure 7.4a shows the detected positions of the patella and of the femur in the SQR scan of the knee. The practitioner positions the probe, confirming that the markers are positioned correctly, and acquires a single US image for each scan. While performing this procedure, the practitioner also specifies the scan depth and gain parameters that are stored by GAJA. Note that these parameters should be tuned for each scan and patient.

Self-acquisition step: Self-acquisition requires the patient to complete a set of tasks. First, the target joint to scan is selected

[2] A scan is a specific view of a body part obtained by positioning the probe in a consistent way.



Position the probe to acquire an image suitable for diagnosis.
Make sure, before clicking the acquisition button, that all markers in the relevant scan are recognized:
- Femur
- Patella

Depth: 50 mm

Gain: 69 %

(a) Setup step by the practitioner (b) Self-acquisition by the patient

Figure 7.4: GAJA two-step image acquisition procedure

from a list, thus loading the probe parameters (**Automated action**). The patient then files a clinical history questionnaire (*e.g.*, does the joint hurt?). Then, a screen containing text indications and images (see Figure 7.2) reminds the patient to perform **Reminded actions** which include adding the gel on the probe and its coarse positioning.

Next, the patient fine-tunes the probe position through a **Guided action**. The guidance is provided with two simultaneous modalities. On the right side of the interface (see Figure 7.4b), visual feed from the probe is overlaid with the bounding boxes of the detected anatomical markers present in the feed (continuous border). The patient has to align the anatomical markers with the target areas (with dashed border) of the same color, which are extracted from the reference image collected by the practitioner. For example, in Figure 7.4b the solid boxes represent the current position of the patella (orange) and femur (purple), while the dashed rectangles represent their position in the reference image.

The left frame provides symbolic indicators of the quality of the current positioning, rendered as sliders, each corresponding to a positioning parameter. Icons at the beginning and at the end of the slider indicate the range of the movements that govern the corresponding parameter. The sliding indicator is centered on the line when the positioning is correct, and it is displaced laterally if the patient needs to perform alignment corrections with respect to that parameter. In the above example, the upper sliding indicator is displaced slightly to the left, indicating that the probe should be approached to the knee.

Once the probe is correctly positioned (see Section 7.3.2) a message informs the user to hold the probe still for 3 seconds. This was required because we empirically observed that the first acquired images are motion-blurred and requiring the patient to hold the device still mitigates this problem.

We also observe that, although the probe is correctly positioned, it is possible that the acquired image is unsuitable, for example, due to blurriness or lack of gel. However, sending an unsuitable image to the practitioner would result in a delay in the diagnosis process, a loss of time for the practitioner, and in a frustrating user experience for the patient. To mitigate this problem, we adopted two solutions. First, GAJA acquires a set of images (instead of a single one), as this increases the chances that at least one of them is suitable. Second, we use a ML model to check if at least one of the collected images is suitable. Thanks to this model, the user can be immediately informed if no image is suitable and can re-acquire the images. Once a set of images is acquired, they are sent to the server where they are stored for the medical practitioner to use in order to formulate a diagnosis. Note that the larger the set of images sent to the medical practitioners, the longer it could take them for finding a suitable one. In order to speed up the diagnosis process, images are ordered according to the suitability as computed by the ML model, so that the images with the highest likelihood of being suitable can be processed first.

7.3.2 Implementation

Architecture

GAJA was implemented as a Windows application and current prototype runs on *Surface GO3*[3], a touchscreen-based portable device.

[3] https://www.microsoft.com/en-us/d/surface-go-3

The device is connected to the portable probe *MicrUS Pro-L40S* manufactured by *Telemed, Lithuania*[4] through a USB-C cable. The application requires bi-directional communication with the ultrasound probe in order to acquire images in real-time and to change the settings (*e.g.*, depth and gain). This was achieved through an SDK made available by the probe manufacturer. Thanks to this solution, the GAJA app can access the US stream of images in real-time, hence making it possible to locally process the images and show the result in real-time.

The data produced by GAJA (questionnaire answers, images, and other metadata, including the detected bounding boxes, and acquisition time) are transmitted to a remote server, hosted at the hospital, which stores them. The server also hosts a web app that the practitioner can use to visualize the data acquired by various patients through GAJA and to provide the diagnoses.

Implementation of machine learning models

GAJA uses two machine learning models: one to detect the anatomical markers, and the other for classifying images suitability.

In order to implement the former model we trained a `YOLO V5`[201] architecture that provides a `nano` version specifically designed to require low memory and provide fast computation also on low-performance devices. Our preliminary results were obtained on a dataset composed of both the Knee SQR and a set of 100 images collected using the portable probe, where the same practitioner annotated the reperees and show a mean Average Precision at 0.5 IoU (mAP@0.5) of 0.986 and 0.922 for the patella and femur, respectively. The model was then exported in the *onnx* format to be used in GAJA. The model processing time on the portable device is about 150 milliseconds. Considering the other computations that are required for each frame (e.g., drawing the bounding boxes, acquiring the frame) GAJA is able to process approximately 4 images per second.

The detection model returns, for each processed frame, the bound-

[4] `https://www.telemedultrasound.com/micrus-pro`

ing boxes of the detected anatomical markers. Since the model can recognize each anatomical marker more than once in each frame, we only consider the prediction for each class with the highest confidence.

The bounding boxes are then displayed as an overlay over the US frame stream. In order to smooth the movements of the bounding boxes as they appear to the user, we adopted a moving average that considers the current and the two previous frames.

Preliminary results suggest that the features that most impact image suitability are the horizontal position of the center of the patella bounding box and the vertical position of the center of the femur bounding box. Hence, for each processed frame, the procedure computes the horizontal distance between the centers of the patella bounding boxes of the current frame and of the reference image. If the distance is smaller than a given threshold, the patella is considered in the correct position. Similarly, GAJA detects if the femur is in the correct position by considering the vertical distance. If both the patella and the femur are in the correct position, then the probe is correctly positioned.

The latter model (classification of image suitability) was implemented as a convolutional neural network based on **InceptionV3**[242] and was trained on the Knee SQR dataset, using all the initially discarded non-suitable images. Our preliminary results show an average `F1-score` of 0.85. In this particular task, the processing time is slower as a result of the model complexity. Hence this model does not run in real-time. The model is currently running on the device but we plan to run it on the server in the future.

7.4 Conclusions

In this Chapter we introduce PRACTICE, a healthcare system specifically designed to support hemophilic patients and the medical practitioners assisting them. The system was also designed with a third main actor in mind: the data scientist who uses the collected data to train new ML models. This required defining medical procedures and technical solutions for the acquisition, annotation, and storage of US media.

All three PRACTICE components are currently being used: CADET supports the practitioner during visits, and ATOM makes it possible to assign annotation tasks to practitioners. The third component, GAJA, is currently available for ten patients who use it to acquire weekly images of the knee recess.

The main limitation of this contribution is the lack of formal system validation, in particular from the point of view of its usability by the patient for the GAJA app, and for the practitioners in the CADET web app. Another limitation of the current approach is the lack of a complete integration with the Italian healthcare system: While our solution facilitates and automates some of the document production, it still needs to be manually inserted into the healthcare IT system. For example, while we can register in our system the need for a follow-up visit, this cannot be automatically reflected in the hospital booking system.

Finally, the GAJA app has several simplifications. Currently, the only supported view is the knee SQR, while we analyzed, by talking to practitioners, that in rarer cases the distension might not be visible in this view and other views (e.g., lateral, medial) might be required. Another limitation is the support for other joints; in fact, as described in Chapter 2, hemophilia affects mainly the knee, elbow, and ankle. However, the latter two are more complicated to acquire, especially in the ankle, where small movements of the probe can result in large changes in the image that would make it unsuitable, and the development of the guiding system is still in the prototyping phase. Moreover, GAJA only allows the acquisition of static images, while recent discoveries indicate that it is necessary to work toward short videos where the recess is squeezed and released. This allows analysis of the movement of the particulate within it, which facilitates the identification of the presence of blood.

8

Conclusions and Future Work

8.1 Summary

In this thesis, we addressed several challenges posed by the limitations of available data in medical imaging, particularly in rare pathologies like hemophilia and emerging conditions such as MPOX.

First, we tackled the issue of data scarcity by employing transfer learning to leverage pre-trained models on larger datasets, thereby improving generalization. Additionally, we applied multi-task learning to extract knowledge across different tasks, further enhancing model performance despite limited data.

Secondly, class imbalance, a significant issue caused by the reduced frequency of hemophilic swelling and limited access to medical facilities, was addressed using an anomaly detection framework. This allowed for training the model solely on *Non-distended* images, with weak supervision provided through bounding boxes, helping to improve sensitivity in the detection of rare conditions.

To mitigate the challenges of domain-shift, we proposed a new domain adaptation technique, which uses cross-reconstruction as auxiliary task to adapt the model to the new domain. This was also applied in similar US images, such as the knee and elbow joints, without requiring large labeled datasets for each region. Through these methods, we demonstrated that it is possible to improve model performance and adaptability in the context of highly imbalanced and scarce medical datasets, paving the way for more effective deep learning applications in rare pathology detection.

Finally, as the ultimate goal of my Ph.D. was to provide a supportive tool for the management of hemophilia, both for patients and practitioners, we integrated the previously discussed contributions into a unified system: PRACTICE. This system comprises three key elements designed to streamline hemophilia care. First, GAJA is an application that offers an automated guiding system to help patients acquire ultrasound images of their joints independently. Second, CADET employs AI methods to help clinicians diagnose hemarthrosis. Lastly, ATOM provides a platform for clinicians to efficiently annotate ultrasound images. The PRACTICE system is currently in use at the Policlinico of Milano, while ten patients are currently undergoing a preliminary test on the self-collection of US images with the GAJA app.

8.2 Discussion and limitations

In this thesis, we faced several challenges and introduced computer vision solutions to address problems related to medical data. The contribution is twofold: first, we proposed solutions that can be generalized to various computer vision problems in the medical domain and possibly also in other domains; second, we specifically addressed the unmet needs of current hemophilia management practices. In this section, we will briefly discuss the results obtained, their implications for the management of hemophilia, and the limitations of the proposed approaches.

8.2.1 Impact

The direct impact of our methods on current hemophilia treatment practices can be divided into three aspects: improved patient outcomes, enhanced practitioner efficiency, and cost savings for the healthcare system.

Concerning patients, the GAJA app, which allows self-collection, would reduce the need for frequent hospital visits, reducing long waiting times, often caused by the limited availability of highly specialized practitioners. It would also allow every patient to make routine visits that may lead to the early detection of blood effusion in the absence of pain. This will also lead to fewer cases of untreated joints, with a reduced risk of permanent damage and, at the same time, to a reduction in the risk of overtreatment and a reduction in the risks associated with excessive treatment, such as inhibitor development, infections, and thrombosis [243]. This self-collection capability not only improves accessibility, but also ensures timely monitoring and intervention, potentially leading to better health outcomes and an increased quality of life for patients with hemophilia.

For practitioners, the CADET web app is designed to guide the examination procedure, potentially improving diagnostic accuracy. This also significantly reduces the time needed to complete the examination, as no specific tool was previously available, which required doctors to manually write the diagnoses for different joints on paper, calculate the HEAD-US score, and then copy everything into the healthcare system. Our system enables a more direct completion of selected fields, thereby reducing the possibility of human errors and speeding up the whole process. As an unintended benefit, the design of the tools and the analysis of the requirements for the various elements of this thesis led to the definition of a standardized procedure that was still missing for the early detection of effusions using US evaluation. This work is currently being formalized and will soon be submitted by our research group. Another positive impact is on the training of less experienced practitioners in the analysis of ultrasound images. The ability to view and analyze multiple sessions, along with annotations and diagnoses made by more experienced practitioners directly on the web app, makes CADET, in conjunction with ATOM, an effective training platform. This allows practitioners to compare their diagnoses and discuss any differences.

Finally, the implementation of PRACTICE in actual medical facilities could impact both direct and indirect costs. For direct costs, reducing overtreatment would minimize the expenses associated with replacement drugs, which can add up to more than $220,000$ yearly euros per patient [244]. It would also lower hospitalization costs by reducing the need for surgeries to mitigate permanent joint damage. Furthermore, the remote acquisition app would reduce the number of in-hospital visits, allowing practitioners to dedicate more time to the most urgent patients. For indirect costs, improved joint health in patients would reduce expenses such as transportation costs, follow-up visits, and impact on educational and professional activities due to mobility limitations or frequent absences for medical appointments. This would also limit the workload of the healthcare system by reducing the number of patients who require hospital care.

Although an accurate health technology impact assessment is currently being conducted at the University of Milan, a preliminary evaluation on economic sustainability can be made. Without taking into consideration the costs of developing and reseaching, the cost per user is limited by the hardware costs. In this specific scenario, patients receive a tablet of around €300 and a portable probe that costs around €$4,000$, which will be potentially used by patients for many years. If we consider a single injection of replacement drugs at around €$2,000$, avoiding two injections could pay for all hardware costs. Furthermore, this does not consider the costs of hospital visits, transportation, etc.

8.2.2 Applicability limitations

Despite the contributions proposed in this thesis, it is important to reflect on the limitations to better understand the impact and to define future research directions.

As the primary goal in hemophilia management is the identification of blood effusion, the extremely limited available data and the complexity of distinguishing between synovial fluid and blood effusion on US images made it impractical to focus directly on this problem. Instead, we directed our efforts to the simpler task of detecting recess distension, which serves as a necessary condition to identify the presence of blood within the joint.

Furthermore, the research problem is evolving, even from a medical perspective. Through the studies conducted in these years, we have realized that subjective annotation is highly dependent on the practitioner's skill level. In addition, the interpretation of the US image by the practitioner is influenced by their knowledge of the patient's medical history. In the last version of the dataset, we acquired annotations from three different practitioners.These results allowed practitioners to discuss and better define the problem. Another requirement that has recently changed is to classify the distention into four classes (instead of two, as done in this thesis). The rationale is that, since distension can occur in various levels of severity, the anatomical characteristics can vary greatly within a single class considering only two classes. We are therefore moving toward a 4-class classification approach, where the levels are absent, minimum, mild, and severe, where the level of the distension is defined by precise anatomical measures.

Another key limitation with respect to the feasibility of the PRACTICE system is that all works presented in Chapters 4, 5, and 6 are based on images collected with a single hospital US probe, the quality of which is higher than the portable probes that are used to collect images in the POC. More evaluations should be performed to evaluate the ability to generalize from the hospital dataset to the portable one. Fortunately, this is supported by the studies on domain adaptation that show the potential of adapting the model to similar domains.

The dataset used in this evaluation presents several limitations that could affect the generalizability of the results. The majority of patients are Italians and are currently being treated with replacement drugs. This condition introduces potential bias, given that the anatomical and medical conditions of the patients are influenced by ongoing treatments. Furthermore, all US images were collected using the same machine, which may limit the diversity of imaging quality and techniques. This uniformity may not reflect the variability in the equipment used in other medical settings. In regions where replacement drugs are less available, the conditions of the patient's joints could be worse, and as a result, the appearance of recesses could differ, further complicating the applicability of these findings to diverse populations.

An aspect that significantly limited the progress of this research was the extensive annotation time required, and indeed the annotation process took more than a year. This thesis aimed to reduce the reliance on a heavily annotated dataset, which would be ideal for training standard deep learning models.

Another limitation is that the dataset produced could not be shared outside the university due to privacy restrictions, which restricted other interested groups from conducting research that could improve performance, validate our findings, or provide valuable comparisons. This lack of access also limits potential collaboration, which is crucial for advancing the field and ensuring the generalizability and robustness of our methods.

Finally, to improve trustworthiness, it is essential to provide practitioners with explanations of how the model arrived at specific decisions. This concept is commonly referred to as eXplainable AI (XAI). In Chapter 3, we demonstrate how this can assist during diagnosis. Currently, the CADET system lacks implementation of this feature, which will be indispensable to provide practitioners with the necessary support.

8.3 Future works

8.3.1 Blood effusion detection

As future work, we are beginning to investigate segmentation techniques to extract the recess area in the US images. This approach will enable precise measurements in centimeters, allowing us to manually extract key features such as the perimeter, area, and horizontal segment that bisects the recess at half its height. These measurements are important because distension is currently subjectively assessed based on these measurements.

As discussed in Section 8.2.2, one main limitation on the applicability of our approach is the inability of the models to differentiate between synovial and blood effusion, the latter being the actual manifestation that requires treatment. Knowing that a possible approach to identifying the presence of blood is by looking at how the particolate moves inside the recess after a dynamic manover: the probe must be squeezed alternately, so as to compress and relax the recess. In the presence of synovial liquid, the liquid appears without swirling movements. In contrast, in the presence of blood, small spekles move within the recess in a cahotic pattern (similar to the effect of a snow globe) [26]. Given this context, we plan to adopt the segmentation technique to isolate the recess area to analyze the content of the recess, without the distraction caused by the typical noise of the US on the muscles and tissues surrounding it. It is important to note that, as mentioned previously, the data acquisition process is costly and time consuming for the practitioners. Fully supervised approaches might not be ideal in this scenario, and supported by the results obtained with weak supervision, we intend to extend LoRIS, giving more focus on the segmentation performances. In addition, temporal enforcement could be adopted to obtain consistent segmentation in videos [245].

The acquisition of images and videos of blood effusions is highly complex. Although the ground-truth could be assessed directly by puncturing the recess and analizing the fluid, this is an invasive approach that cannot be easily performed on all patients. In addition, visible effusions presented in the center occur only a couple of times each year. This is because to collect images of this condition, patients must have ongoing bleeding and visit our facility, whereas they typically go to the emergency room, where our system is not integrated. This highlights the importance of a multicenter study in collecting a sufficient amount of data. However, the total number of images will remain small, underscoring the need for machine learning techniques that can work effectively with limited data. To facilitate the procedure, we defined a standardized protocol with the practitioners to collect and annotate the images without puncturing the patients. When there is a suspicion of bleeding, the patient is treated, and after 2 to 4 days, if swelling reduces, it indicates that the content was blood. In contrast, if swelling persists, it was caused by something else. This allows us to accurately identify which episodes and their corresponding images were of a blood effusion. Once these results are obtained, we can start analyzing the dynamic patterns that occur inside the recess by looking, for example, at the optical flow or using anomaly detection in videos to find unusual behavior, as described above.

8.3.2 Improved CAD system

To extend the work done on the PRACTICE system there are several research directions that need to be addressed. The first would be to integrate the system that allows analysis and detection of distention in other joints, such as the elbow and ankle, together with their different views. This can be achieved through the adoption of domain adaptation techniques when really similar (e.g. knee to elbow), while it will probably require to collect and annoate new data for completely different anathomical structures such as the ones of the ankles.

In future work, it will be essential to improve model calibration to improve the reliability of the predictions, particularly when using them as an index of the presence of a pathology. A well-calibrated model should exhibit a strong inverse relationship between prediction confidence and error rate, which means that higher certainty should correlate with higher accuracy, ultimately supporting more informed decision making. This is important for clinical applications, as overconfident but incorrect predictions could lead to under and overtreatment.

Furthermore, GAJA has been currently evaluated with non-hemofilic patients, in a group that is statistically more inclined toward the use of technological devices. In fact, we are conducting a user study, with the aim of assessing the ability of the system to guide hemophilic patients to collect reliable US scans. Currently, 9 patients out of the target 13 are participating in the tests. 3 of the patients have completed the 19-week study in which they were asked to collect an image each week. A longitudinal study will also measure the extent to which GAJA can be used by patients over a long period, as previous work [41] revealed that it can be challenging for patients to remember how to use the system. We conjecture that, since GAJA adopts the *Automate-Guide-Remind* design principle, it will substantially mitigate this problem.

To deal with the limit imposed by the lack of portable-probe annotated images, we plan to adopt ReC-TTT to assess the performances of a model trained with only images generated by high-end probes, on low-end probes. Similarly, the technique will require to be tuned to achieve sufficient performances on the knee-to-elbow adaptation.

Since thrustworthiness and transparency are essential, especially in the medical domain, we plan to adopt XAI techniques to better identify the models' choices. This aims to increase the reliability of the predictions and suggest which parts of the images or videos should be relevant to the practitioners to make the diagnosis. It will also be essential to address the technological validation process and legal approval of the developed tool before it can be introduced as a product for patients. This involves ensuring that the technology meets regulatory standards and demonstrates safety and efficacy through rigorous testing and conducting comprehensive clinical trials.

8.3.3 Dataset publication

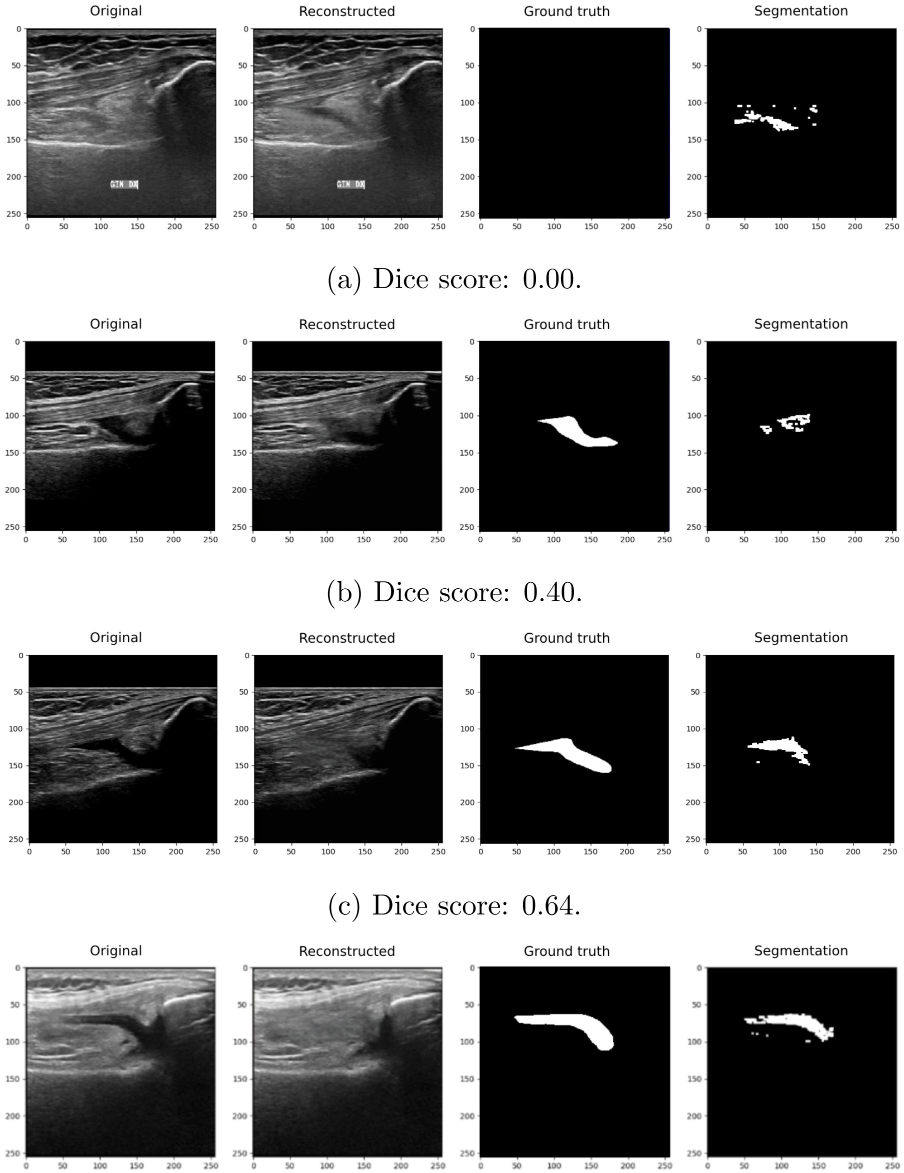
We are discussing with the ethics committee the procedures for data collection, annotation and obtaining informed patient consent for the public sharing of anonymized data, to ensure ethical compliance. The ground truth annotations of distension and blood effusion will be based on the diagnosis generated during visits and together with the knowledge of the patient's history, and we will provide annotations (based on individual images) made by different practitioners. A future research direction will be to explore how leveraging the concordance and discordance between them can improve the approximation of the ground truth and lead to a more calibrated model. An intra and inter practitioners assessment will be conducted.

A

Visualization of LoRIS results

In this section, we provide examples of the qualitative evaluation of LoRIS results (Figure A.1).

Note that false positives can occur when the model fails in the reconstruction of nomral recesses (see Figure A.1a), in this case the dice score is 0, therefore lowering the average performance of the model. In the second example (Figure A.1b) the model correctly recontructed the image, but the reconstructed recess still appears partially swallen and therefore the distance algorithm is not capable of identifying the whole anomalous area. Finally, the last two examples (Figures A.1c and A.1d) show a good reconstruction and anomaly detection, leading to higher dice scores (0.64 and 0.77).



(a) Dice score: 0.00.

(b) Dice score: 0.40.

(c) Dice score: 0.64.

(d) Dice score: 0.77.

Figure A.1: Qualitative evaluation of LoRIS reconstructions and segmentations.

B

ReC-TTT extended results

For CIFAR-100C, TinyImageNet-C and VisDA our model was compared with the same state-of-the-art approaches except TTT++ where the results were not reproducible nor available: ResNet50 [178], PTBN [135], TENT [136], TIPI [137], ClusT3 [132] and *NC-TTT*[143]. As per previous experiments TTA methods were evaluated on the same pre-trained ResNet50, while TTT approaches were trained using the same ResNet50 base architecture and the same training strategy.

B.1 VisDA

Table B.1 reports the detailed results on the VisDA dataset. *ReC-TTT* outperforms most approaches on average, with a notable increase compared to the ResNet50 baseline without adaptation (+25.81). On *train* → *val* and *train* → *test*, *NC-TTT* performs better than *ReC-TTT* ($\approx +1\%$ on average). Moreover, the results demonstrate that TTT methods show greater robustness on complex datasets, such as VisDA, compared to methods like Source, PTBN, and TENT, which are more competitive on the CIFAR datasets. This performance difference may be attributed to the reconstruction task's ability to capture more generalizable features, while simpler approaches struggle to detect more subtle domain shifts.

Table B.1: Performance comparison with state-of-the-art on VisDA dataset (%).

	VisDA *train* → *val*	VisDA *train* → *test*	Average
ResNet50	35.01	36.58	35.80
PTBN	54.53	53.63	54.08
TENT	58.13	57.04	57.59
TIPI	60.22	62.26	61.24
ClusT3	60.89	61.33	61.11
NC-TTT	**62.49**	**62.57**	**62.53**
ReC-TTT	62.06	61.12	61.59

B.2 CIFAR-100C

Table B.2 shows in detail the results and the comparison with state-of-the-art approaches on all the perturbations of CIFAR-100C. *ReC-TTT* the best results, demonstrating a 30% increase in AUROC after adaptation compared to the baseline. This improvement surpasses the most recent state-of-the-art approaches as ClusT3 and *NC-TTT* by 3%.

B.2.1 Number of adaptation iterations

Similarly to what was identified in previous studies [139, 132, 143] and was confirmed for CIFAR-10C, also in the case of CIFAR-100C the best results are obtained after 20 adaptation iterations, while for some perturbation the same results can be obtained also with less interaction, after 20 the results tend to remain invariant for all the different perturbations. Figure B.1 shows for all the corruption of CIFAR-100C the results obtained at different iterations.

Corruption Type	**ResNet50**	**PTBN**	**TENT**	**TIPI**	**ClusT3**	***NC-TTT***	***ReC-TTT***
Gaussian Noise	13.23	42.30	51.35	48.88	**52.79**	46.03	48.12
Shot Noise	15.46	43.30	52.63	50.61	**52.91**	47.04	50.43
Impulse Noise	7.89	37.41	45.39	43.80	**45.54**	41.53	45.29
Defocus Blur	27.36	67.46	69.44	68.72	66.66	67.00	**71.21**
Glass Blur	21.18	46.44	**51.01**	50.93	50.76	48.08	49.94
Motion Blur	38.18	64.21	67.27	66.63	62.92	64.31	**68.86**
Zoom Blur	32.81	66.68	69.33	68.84	65.42	66.24	**69.91**
Snow	44.85	55.52	**60.47**	59.51	56.65	58.70	60.21
Frost	31.56	54.76	58.35	57.90	56.91	58.55	**60.16**
Fog	32.79	56.77	**62.29**	61.12	53.95	57.73	62.22
Brightness	66.13	68.97	71.40	71.00	66.78	71.36	**73.47**
Contrast	11.87	63.47	65.63	65.17	56.46	61.53	**67.06**
Elastic Transform	48.87	57.93	60.07	59.94	59.07	60.25	**62.37**
Pixelate	26.70	59.75	**64.06**	63.56	62.26	61.17	63.61
JPEG Compression	48.88	52.45	57.84	57.79	**59.34**	55.69	57.05
Average	31.19	55.83	60.44	59.63	57.89	57.68	**60.66**

Table B.2: Performance comparison with state-of-the-art on CIFAR-100C perturbations (%).

B.3 TinyImagenet-C

Table B.3 reports the results obtained on TinyImagenet-C, a dataset of 10.000 images with the same 15 corruptions described for CIFAR10-C and CIFAR100-C, but with 200 classes. *ReC-TTT* outperforms all the other methods also on this dataset, with a 2.46% improvement compared to *NC-TTT*, the second-best-performing model.

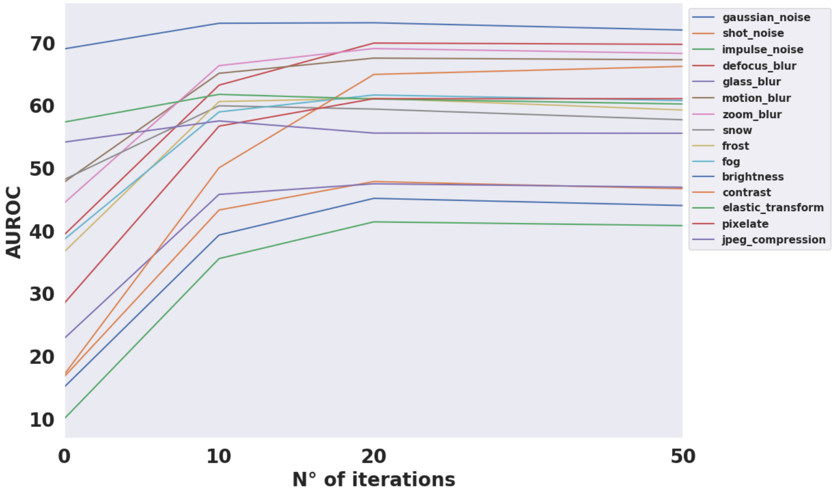


Figure B.1: Performance (AUROC) reached by our method with different numbers of adaptation iterations on CIFAR-100C.

Corruption Type	**ResNet50**	**PTBN**	**TENT**	**TIPI**	**ClusT3**	***NC-TTT***	***ReC-TTT***
Gaussian Noise	13.20	30.46	31.03	32.22	32.65	31.92	**34.87**
Shot Noise	16.28	32.26	33.07	34.27	34.72	34.47	**36.60**
Impulse Noise	7.49	20.80	21.87	23.04	22.78	22.78	**26.09**
Defocus Blur	16.71	33.09	**34.20**	31.98	29.08	25.28	31.09
Glass Blur	7.42	15.97	16.88	17.60	16.26	15.67	**19.59**
Motion Blur	27.71	43.09	44.40	43.54	43.92	43.39	**45.55**
Zoom Blur	20.98	39.76	40.89	40.01	41.17	40.46	**42.53**
Snow	31.00	36.94	37.39	38.18	42.97	**43.46**	40.33
Frost	36.28	39.29	40.21	41.43	45.32	**45.51**	44.59
Fog	16.40	31.51	32.52	32.82	**37.85**	37.68	33.08
Brightness	36.48	44.70	45.09	46.39	**51.19**	50.62	48.53
Contrast	2.59	12.22	**12.91**	10.71	2.27	2.27	8.32
Elastic Transform	28.93	39.42	39.83	40.68	41.60	41.47	**44.91**
Pixelate	37.00	47.78	48.50	48.95	37.00	39.31	**52.96**
JPEG Compression	47.04	47.78	40.88	50.21	50.57	50.91	**53.32**
Average	23.03	34.47	35.15	35.47	35.32	35.03	**37.49**

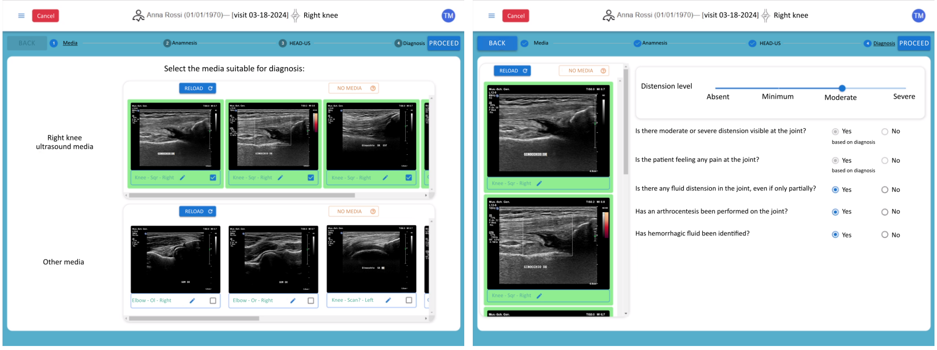
Table B.3: Performance comparison with state-of-the-art on TinyImageNet-C perturbations (%).

C

Details on PRACTICE components

C.1 CADET: Computer-Aided Diagnosis for hEmarThrosis

CADET is a web-based interface that supports clinicians in formulating the diagnosis; it manages both in-presence and remote visits. To design CADET we first analyzed the habitual visit procedure adopted by practitioners without the support of a computer-aided diagnosis system. The physicians used to collect media with the ultrasound probe and then enter the diagnosis of blood effusion in a word processor file, following a template that defines a set of information for each joint [14]. The diagnosis was finally uploaded to the national health system server and, after printing, stored in the patient's physical medical record.



Select the media suitable for diagnosis:

Right knee ultrasound media

Other media

Distension level: Absent, Minimum, Moderate, Severe

Is there moderate or severe distension visible at the joint? ● Yes ○ No
based on diagnosis

Is the patient feeling any pain at the joint? ● Yes ○ No
based on diagnosis

Is there any fluid distension in the joint, even if only partially? ● Yes ○ No

Has an arthrocentesis been performed on the joint? ● Yes ○ No

Has hemorrhagic fluid been identified? ● Yes ○ No

(a) Images selection

(b) Diagnosis questionnaire

Figure C.1: CADET interface

This procedure had several limitations. First, the media and exam data were not linked, making access to the patient's medical history (complete with diagnosis and the media) impractical. This affects practitioners, who need to review the stored data during follow-up visits, and also makes it impossible to use the data for the training of ML models. Second, some operations required the practitioner's intervention although, in principle, they could be automated. This included, for example, the creation of the diagnosis on the word processor. Finally, no CAD system was implemented and remote visits were not possible.

We initially designed a first CADET prototype in which the practitioner could use the web app to automatically acquire media from the ultrasound probe. This solution was designed with the idea that the practitioner could quickly switch from CADET to the probe. However, due to technical limitations of the ultrasound probe (no SDK is available), this was not possible. Therefore, we designed a solution in which the practitioner first acquires media using the ultrasound probe and then interacts with CADET to formulate the diagnosis. The practitioner first completes an initial general medical history through a guided questionnaire and then selects the joint, one at a time. The diagnosis of each joint is divided into four steps: media selection (Figure C.1a), joint-specific history, a questionnaire related to the standard HEAD-US procedure [14], and a guided questionnaire for the diagnosis of blood effusion (Figure C.1b).

After completing the process for each joint, the practitioner can access the final diagnosis that follows the same format as the word processor template. This report can then be uploaded to the national health system server and possibly printed for physical storage.

The remote diagnosis procedure is similar, with the main difference that some information is already available (media and history).

CADET adopts two main solutions to support the practitioner. First, it implements a knowledge-based system to guide the practitioner in diagnosis formulation. This solution was first designed in terms of a decision tree in which each node is a Boolean condition and each leaf is the join-specific medical report. CADET implements this decision tree through a questionnaire (see Figure C.1b) in which some answers are automatically provided based on the data inserted in the previous steps (*e.g.*, whether the patient has pain) and the remaining are provided by the practitioner. The second solution adopted to support the diagnosis is to automatically detect recess distention [1], which is a necessary condition for blood effusion. Taking into account the media available for a given joint, the system suggests a distention value on a scale of four possible alternatives (see Figure C.1b). The practitioner can then decide to accept the suggested value or to change it.

Several solutions were also adopted to speed up the process. First, CADET automatically pre-selects the media obtained from a visit based on a ML solution that identifies, for each media, the scan, the joint name, and its laterality. For each joint, some data are precomputed on the basis of previous visits and the patient's medical history. For example, for each joint, the practitioner has to specify whether there is a prosthesis. If the practitioner specifies that there is one during a visit, the system automatically loads the same value during the following visits. Finally, CADET automatically generates the diagnosis file that can be uploaded to the national system.

C.2 ATOM: Annotation Task Orchestrator Module

Deep learning algorithms rely on large datasets to effectively learn to generalize patterns of various pathological conditions or to identify areas of interest. However, a public dataset of US media is not available for the medical domain considered. Therefore, we created a new dataset by collecting US media from hemophilic patients. Data was collected by expert practitioners during hospital visits, using a high-end ultrasound device. Since media is acquired during regular visits, the overall procedure was designed to avoid additional workload for the practitioners and inconvenience to patients.

One problem that emerged during the creation of the dataset is related to the fact that US imaging is highly dependent on the operator and has a high inter-patient variability hence making the acquired data highly heterogeneous, a factor that can negatively impact the training of the machine learning models. To mitigate this problem, we defined an acquisition protocol based on the following principles [1, 26].

- **Inclusion criteria**. The media of patients with significantly different characteristics (at the level of musculoskeletal US imaging) are excluded from the dataset. For example, children and patients with prostheses are excluded.

- **Standardization**. By adopting well-established procedures in the medical literature and practice, we defined a standard procedure for image acquisition. This includes, for example, the set of joints to consider and, for each of them, the set of scans [1].

- **Parameters definition**. When acquiring an ultrasound media, the practitioner can tune several settings (e.g. power and

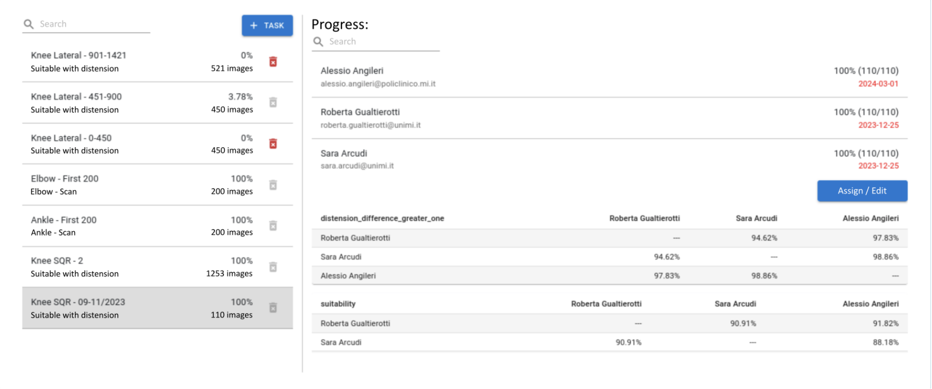
[1] A scan defines the probe position and, consequently, which anatomical targets are framed in the US image.

frequency). We selected a fixed value for most of these settings, leaving the practitioner with the ability to select only a few parameters whose value has to be defined specifically for each patient (*e.g.*, the “depth” value).

After dataset acquisition, we defined a set of tools and practices for data annotation. The guiding objectives were to reduce the annotation time and errors. To achieve these objectives, we initially developed an ad-hoc annotation tool. However, we then realized that the research activities frequently required the creation of new annotation tasks. For this reason, we designed the ATOM (Annotation Task Orchestrator Module) system that allows the administrator to quickly create a new annotation task by specifying the following data.

- The set of media from the dataset.
- The annotation tool, a third-party application. For example, for some annotation tasks, we used *Label Studio* [246] that, for privacy reasons, we configured to run on our server. These tools automatically transmit the annotations to PRACTICE, which stores them.
- The type of annotation (*e.g.*, the set of classes).
- The set of annotators (*i.e.*, practitioners).

The system was designed to interact with any compatible annotation tool, including those for creating image class annotations, bounding boxes, and segmentations. In addition to creating annotation tasks, ATOM also provides two main functions. One function is designed for annotators, who can access the list of tasks assigned to them and run the annotation tool. The other function is designed for the administrator to monitor the completion of the annotation tasks and to check the inconsistencies among the annotators. Specifically, for each annotation task, the administrator can define one or more equality functions. Then ATOM uses these functions to create a confusion matrix that shows, for each pair of annotators, the percentage



Knee Lateral - 901-1421 Suitable with distension	0% 521 images
Knee Lateral - 451-900 Suitable with distension	3.78% 450 images
Knee Lateral - 0-450 Suitable with distension	0% 450 images
Elbow - First 200 Elbow - Scan	100% 200 images
Ankle - First 200 Ankle - Scan	100% 200 images
Knee SQR - 2 Suitable with distension	100% 1253 images
Knee SQR - 09-11/2023 Suitable with distension	100% 110 images

Alessio Angileri alessio.angileri@policlinico.mi.it	100% (110/110) 2024-03-01
Roberta Gualtierotti roberta.gualtierotti@unimi.it	100% (110/110) 2023-12-25
Sara Arcudi sara.arcudi@unimi.it	100% (110/110) 2023-12-25

distension_difference_greater_one	Roberta Gualtierotti	Sara Arcudi	Alessio Angileri
Roberta Gualtierotti	—	94.62%	97.83%
Sara Arcudi	94.62%	—	98.86%
Alessio Angileri	97.83%	98.86%	—

suitability	Roberta Gualtierotti	Sara Arcudi	Alessio Angileri
Roberta Gualtierotti	—	90.91%	91.82%
Sara Arcudi	90.91%	—	88.18%

Figure C.2: ATOM admin example screen

of images (among those annotated by both annotators) that have the same annotation (according to the equality functions). Figure C.2 shows an example screen of the admin panel, on the left we can see the list of active tasks, with the progress. On the right a detailed view of a single selected task, where the progress is divided among the different practitioners and the tables on the bottom report statistics of the currently annotated data for various equality functions.

D

Joint ultrasound views

Table D.1: All the views available in the HEAD-US protocol. *Indicates the most appropriate view for identifying a blood effusion.

Joint	View	Acronym
Ankle	Lateral transverse scan of the sinus tarsi	**ST***
	Anterior longitudinal scan over the tibio-talar joint	TTL
	Posterior longitudinal scan of the tibio-talar joint	TTP
	Anterior transverse scan over the talar dome	TTT
Elbow	Anterior transverse scan of distal humeral epiphysis	DHE
	Posterior longitudinal scan over the olecranon recess	**OLR***
	Anterior longitudinal scan over the radio-humeral joint	RHJ
	Anterior longitudinal scan over the ulno-humeral joint	UHJ
Knee	Anterior longitudinal scan of over the subquadricipital recess	**SQR***
	Medial transverse scan of the parapatellar recess	MED
	Lateral transverse scan of the parapatellar recess	LAT
	Anterior cranial scan of the femoral trochlea	FEM

E

Glossary of Medical Terms

- **Anechoic**: Referring to a structure that does not produce echoes on an ultrasound, appearing completely black; typically indicates the presence of fluid.

- **Blood clot**: A mass formed by platelets and proteins in the blood that helps stop bleeding by sealing wounds in blood vessels.

- **Blood effusion**: The accumulation of blood in a body cavity, often due to trauma, injury, or a medical condition.

- **Hemophilia**: A genetic disorder that impairs the body's ability to make blood clots, leading to prolonged bleeding.

- **Isoechoic**: Describes a tissue or structure that has a similar echogenicity (brightness) to surrounding tissues on an ultrasound, making it difficult to differentiate from them.

- **Musculoskeletal**: Pertaining to the muscles and skeleton, encompassing structures such as bones, joints, ligaments, and muscles.

- **Nonreplacement drugs**: Medications used to manage symptoms or conditions without replacing or supplementing deficient substances in the body.

- **Profilaxis**: A preventive treatment or procedure aimed at reducing the risk of disease or complications.

- **Replacement drugs**: Medications that substitute or supplement deficient hormones, enzymes, or other substances in the body.

- **Subclinical bleeding**: Minor bleeding that does not produce noticeable symptoms and may not be detected without specific tests.

- **Synovial hyperplasia**: An increase in the number of cells in the synovial membrane, often associated with inflammatory conditions and joint diseases.

- **Synovial effusion**: The accumulation of excess synovial fluid in a joint space, often resulting from injury, inflammation, or underlying conditions.

F

Code availability

The code made available can be found at the following links:

- Chapter 3: `https://zenodo.org/records/7981159`
- Chapter 5: `https://github.com/warpcut/LoRIS`
- Chapter 6: `https://github.com/warpcut/ReC-TTT`

Acronyms

- **AD**: Anomaly Detection
- **AI**: Artificial Intelligence
- **ANOVA**: ANalysis Of VAriance
- **API**: Application Programming Interface
- **CAM**: Class Activation Maps
- **CAD**: Computer Aided Diagnosis
- **CDC**: Centers for Disease Control and Prevention
- **CI**: Confidence Interval
- **CNN**: Convolutional Neural Network
- **CT**: Computed Tomography
- **DB**: DataBase
- **DG**: Domain Generalization
- **DL**: Deep Learning
- **DD**: Directional Difference
- **FC**: Fully Connected

- **FN**: False Negative
- **FP**: False Positive
- **GAN**: Generative Adversarial Network
- **GMS**: Gradient Magnitude Similarity
- **HEAD-US**: Hemophilia Early Arthropathy Detection with UltraSound
- **IoU**: Intersection over Union
- **JADE**: Joint Tissue Activity and Damage Examination
- **LR**: Learning Rate
- **mAP**: Mean Average Precision
- **ML**: Machine Learning
- **MPOX**: Monkeypox
- **MRI**: Magnetic Resonance Imaging
- **MTL**: Multi-Task Learning
- **OLR**: OLecranic Recess
- **POC**: Point-Of-Care
- **PRACTICE**: Pilot on Remote AutomatiC ulTrasound scan analysIs for hemophiliC patiEnts
- **REST**: Representational State Transfer
- **SDK**: Software Development Kit
- **SQR**: Sub-Quadricipital Recess
- **SSIM**: Structural Similarity Index Measure

- **TTA**: Test-Time Adaptation
- **TL**: Transfer Learning
- **TN**: True Negative
- **TP**: True Positive
- **TTT**: Test-Time Training
- **UAD**: Unsupervised Anomaly Detection
- **US**: Ultra-Sound
- **WHO**: World Health Organisation
- **WSL**: Weakly-Supervised Learning
- **XAI**: eXplainable Artificial Intelligence

Bibliography

[1] M. Colussi, G. Civitarese, D. Ahmetovic, C. Bettini, R. Gualtierotti, F. Peyvandi, and S. Mascetti, "Ultrasound detection of subquadricipital recess distension," *Intelligent Systems with Applications*, vol. 17, p. 200183, 2023.

[2] M. G. Campana, M. Colussi, F. Delmastro, S. Mascetti, and E. Pagani, "A transfer learning and explainable solution to detect mpox from smartphones images," *Pervasive and Mobile Computing*, vol. 98, p. 101874, 2024.

[3] R. Gualtierotti, A. Giachi, C. Suffritti, L. Bedogni, F. Franco, F. Poggi, S. Mascetti, M. Colussi, D. Ahmetovic, V. Begnozzi, *et al.*, "Optimizing long-term joint health in the treatment of hemophilia," *Expert Review of Hematology*, pp. 1–10, 2024.

[4] M. Colussi, S. Mascetti, D. Ahmetovic, G. Civitarese, M. Cacciatori, F. Peyvandi, R. Gualtierotti, S. Arcudi, and C. Bettini, "Gaja-guided self-acquisition of joint ultrasound images," in *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 132–141, Springer Nature Switzerland, 2023.

[5] D. Ahmetovic, A. Angileri, S. Arcudi, C. Bettini, G. Civitarese, M. Colussi, A. Giachi, R. Gualtierotti, S. Mascetti, M. Manzoni, *et al.*, "Insights on the development of practice, a research-oriented healthcare platform," in *2024 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 380–385, IEEE, 2024.

[6] M. Colussi, A. Giachi, R. Gualtierotti, S. Mascetti, M. Manzoni, *et al.*, "Loris: Weakly-supervised anomaly detection for ultrasound image," in *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. ??–??, Springer Nature Switzerland, 2024.

[7] World Health Organization (WHO), "2022 mpox (monkeypox) outbreak: Global trends." `https://worldhealthorg.shinyapps.io/mpx_global/`. Accessed: 2023-02-08.

[8] G. Roosendaal and F. P. J. G. Lafeber, "Blood-induced joint damage in hemophilia: Modern management of hemophilia a to prevent bleeding and arthropathy," *Seminars in thrombosis and hemostasis*, vol. 29, no. 1, pp. 37–42, 2003.

[9] A. Srivastava, E. Santagostino, A. Dougall, S. Kitchen, M. Sutherland, S. W. Pipe, M. Carcao, J. Mahlangu, M. V. Ragni, J. Windyga, *et al.*, "Wfh guidelines for the management of hemophilia," *Haemophilia*, vol. 26, pp. 1–158, 2020.

[10] M. W. Hilgartner, "Current treatment of hemophilic arthropathy," *Current opinion in pediatrics*, vol. 14, no. 1, pp. 46–49, 2002.

[11] D. Plut, B. F. Kotnik, I. P. Zupan, D. Kljucevsek, G. Vidmar, Z. Snoj, C. Martinoli, and V. Salapura, "Diagnostic accuracy of haemophilia early arthropathy detection with ultrasound (head-us): a comparative magnetic resonance imaging (mri) study," *Radiology and oncology*, vol. 53, no. 2, pp. 178–186, 2019.

[12] P. N. T. Wells, "Ultrasound imaging," *Physics in medicine & biology*, vol. 51, no. 13, pp. R83–R98, 2006.

[13] F. Joshua, M. Lassere, A. K. Scheel, D. Kane, W. Grassi, P. G. Conaghan, R. J. Wakefield, M.-A. D'Agostino, G. A. Bruyn,

M. Szkudlarek, E. Naredo, W. A. Schmidt, P. Balint, E. Filippucci, M. Backhaus, and A. Iagnocco, “Summary findings of a systematic review of the ultrasound assessment of synovitis,” *Journal of rheumatology*, vol. 34, no. 4, pp. 839–847, 2007.

[14] C. Martinoli, O. D. C. Alberighi, G. Di Minno, E. Graziano, A. C. Molinari, G. Pasta, G. Russo, E. Santagostino, A. Tagliaferri, A. Tagliafico, *et al.*, “Development and definition of a simplified scanning procedure and scoring method for haemophilia early arthropathy detection with ultrasound (head-us),” *Thrombosis and haemostasis*, vol. 109, no. 06, pp. 1170–1179, 2013.

[15] Z. Long, X. Zhang, C. Li, J. Niu, X. Wu, and Z. Li, “Segmentation and classification of knee joint ultrasonic image via deep learning,” *Applied Soft Computing*, vol. 97, p. 106765, 2020.

[16] R. Gualtierotti, L. P. Solimeno, and F. Peyvandi, “Hemophilic arthropathy: current knowledge and future perspectives,” *Journal of Thrombosis and Haemostasis*, vol. 19, no. 9, pp. 2112–2121, 2021.

[17] M. Lewandowska, S. Nasr, and A. D. Shapiro, “Therapeutic and technological advancements in haemophilia care: Quantum leaps forward,” *Haemophilia*, vol. 28, pp. 77–92, 2022.

[18] P. M. Mannucci, “Hemophilia treatment innovation: 50 years of progress and more to come,” *Journal of Thrombosis and Haemostasis*, vol. 21, no. 3, pp. 403–412, 2023.

[19] G. Roosendaal and F. P. Lafeber, “Blood-induced joint damage in hemophilia,” in *Seminars in thrombosis and hemostasis*, vol. 29, pp. 037–042, Copyright© 2003 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New . . . , 2003.

[20] N. Bakeer, S. Dover, P. Babyn, B. M. Feldman, A. von Drygalski, A. S. Doria, D. M. Ignas, A. Abad, C. Bailey, I. Beggs,

et al., “Musculoskeletal ultrasound in hemophilia: results and recommendations from a global survey and consensus meeting,” *Research and practice in thrombosis and haemostasis*, vol. 5, no. 5, p. e12531, 2021.

[21] M. N. D. Di Minno, P. Ambrosino, G. Quintavalle, A. Coppola, A. Tagliaferri, C. Martinoli, and G. F. Rivolta, “Assessment of hemophilic arthropathy by ultrasound: where do we stand?,” in *Seminars in Thrombosis and Hemostasis*, vol. 42, pp. 541–549, Thieme Medical Publishers, 2016.

[22] M. Di Minno, S. Iervolino, E. Soscia, A. Tosetto, A. Coppola, M. Schiavulli, E. Marrone, C. Ruosi, M. Salvatore, and G. Di Minno, “Magnetic resonance imaging and ultrasound evaluation of “healthy” joints in young subjects with severe haemophilia a,” *Haemophilia*, vol. 19, no. 3, pp. e167–e173, 2013.

[23] S. N. Keshava, S. Gibikote, and A. S. Doria, “Imaging evaluation of hemophilia: musculoskeletal approach,” in *Seminars in thrombosis and hemostasis*, vol. 41, pp. 880–893, Thieme Medical Publishers, 2015.

[24] S. Nguyen, X. Lu, Y. Ma, J. Du, E. Chang, and A. Von Drygalski, “Musculoskeletal ultrasound for intra-articular bleed detection: a highly sensitive imaging modality compared with conventional magnetic resonance imaging,” *Journal of Thrombosis and Haemostasis*, vol. 16, no. 3, pp. 490–499, 2018.

[25] A. von Drygalski, R. E. Moore, S. Nguyen, R. F. Barnes, L. M. Volland, T. H. Hughes, J. Du, and E. Y. Chang, “Advanced hemophilic arthropathy: sensitivity of soft tissue discrimination with musculoskeletal ultrasound,” *Journal of Ultrasound in Medicine*, vol. 37, no. 8, pp. 1945–1956, 2018.

[26] R. Gualtierotti, L. P. Solimeno, F. Peyvandi, A. Giachi, S. Arcudi, A. Ciavarella, and S. M. Siboni, “Ultrasound evaluation of

hemophilic arthropathy: a proposal of definitions in a changing landscape," 2024.

[27] H. De la Corte-Rodriguez, E. C. Rodriguez-Merchan, M. T. Alvarez-Roman, M. Martin-Salces, C. Martinoli, and V. Jimenez-Yuste, "The value of head-us system in detecting subclinical abnormalities in joints of patients with hemophilia," *Expert review of Hematology*, vol. 11, no. 3, pp. 253–261, 2018.

[28] L. M. Volland, J. Y. Zhou, R. F. Barnes, R. Kruse-Jarres, B. Steiner, D. V. Quon, C. Bailey, T. H. Hughes, R. E. Moore, E. Y. Chang, *et al.*, "Development and reliability of the joint tissue activity and damage examination for quantitation of structural abnormalities by musculoskeletal ultrasound in hemophilic joints," *Journal of Ultrasound in Medicine*, vol. 38, no. 6, pp. 1569–1581, 2019.

[29] S. P. Grogan and C. A. Mount, "Ultrasound physics and instrumentation," 2021.

[30] R. Kulkarni, "Use of telehealth in the delivery of comprehensive care for patients with haemophilia and other inherited bleeding disorders," *Haemophilia*, vol. 24, no. 1, pp. 33–42, 2018.

[31] Q. Huang, Y. Zheng, M. Lu, and Z. Chi, "Development of a portable 3d ultrasound imaging system for musculoskeletal tissues," *Ultrasonics*, vol. 43, no. 3, pp. 153–163, 2005.

[32] G.-D. Kim, C. Yoon, S.-B. Kye, Y. Lee, J. Kang, Y. Yoo, and T.-K. Song, "A single fpga-based portable ultrasound imaging system for point-of-care applications," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 59, no. 7, pp. 1386–1394, 2012.

[33] G. Corte, S. Bayat, K. Tascilar, L. Valor-Mendez, L. Schuster, J. Knitza, F. Fagni, G. Schett, A. Kleyer, and D. Simon,

"Performance of a handheld ultrasound device to assess articular and periarticular pathologies in patients with inflammatory arthritis," *Diagnostics*, vol. 11, no. 7, p. 1139, 2021.

[34] N. M. Duggan, N. Jowkar, I. W. Ma, S. Schulwolf, L. A. Selame, C. E. Fischetti, T. Kapur, and A. J. Goldsmith, "Novice-performed point-of-care ultrasound for home-based imaging," *Scientific Reports*, vol. 12, no. 1, p. 20461, 2022.

[35] A. T. Chiem, G. W. Lim, A. P. Tabibnia, A. S. Takemoto, D. M. Weingrow, and J. E. Shibata, "Feasibility of patient-performed lung ultrasound self-exams (patient-plus) as a potential approach to telemedicine in heart failure," *ESC Heart Failure*, vol. 8, no. 5, pp. 3997–4006, 2021.

[36] E. Schneider, N. Maimon, A. Hasidim, A. Shnaider, G. Migliozzi, Y. S. Haviv, D. Halpern, B. Abu Ganem, and L. Fuchs, "Can dialysis patients identify and diagnose pulmonary congestion using self-lung ultrasound?," *Journal of Clinical Medicine*, vol. 12, no. 11, p. 3829, 2023.

[37] Y. Baribeau, A. Sharkey, O. Chaudhary, S. Krumm, H. Fatima, F. Mahmood, and R. Matyal, "Handheld point-of-care ultrasound probes: the new generation of pocus," *Journal of cardiothoracic and vascular anesthesia*, vol. 34, no. 11, pp. 3139–3145, 2020.

[38] P. B. McBeth, I. Crawford, M. Blaivas, T. Hamilton, K. Musselwhite, N. Panebianco, L. Melniker, C. G. Ball, L. Gargani, C. Gherdovich, *et al.*, "Simple, almost anywhere, with almost anyone: remote low-cost telementored resuscitative lung ultrasound," *Journal of Trauma and Acute Care Surgery*, vol. 71, no. 6, pp. 1528–1535, 2011.

[39] M. Berlet, T. Vogel, M. Gharba, J. Eichinger, E. Schulz, H. Friess, D. Wilhelm, D. Ostler, M. Kranzfelder, *et al.*, "Emer-

gency telemedicine mobile ultrasounds using a 5g-enabled application: development and usability study," *JMIR Formative Research*, vol. 6, no. 5, p. e36824, 2022.

[40] J. Aznar, S. Pérez-Alenda, M. Jaca, M. García-Dasí, C. Vila, A. Moret, F. Querol, and S. Bonanad, "Home-delivered ultrasound monitoring for home treatment of haemarthrosis in haemophilia a," *Haemophilia*, vol. 21, no. 2, pp. e147–e150, 2015.

[41] P. Aguero, R. F. Barnes, A. Flores, and A. von Drygalski, "Teleguidance for patient self-imaging of hemophilic joints using mobile ultrasound devices: A pilot study," *Journal of Ultrasound in Medicine*, vol. 42, no. 3, pp. 701–712, 2023.

[42] H. Culbertson, J. M. Walker, M. Raitor, A. M. Okamura, and P. J. Stolka, "Plane assist: the influence of haptics on ultrasound-based needle guidance," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*, pp. 370–377, Springer, 2016.

[43] S.-Y. Sun, M. Gilbertson, and B. W. Anthony, "Computer-guided ultrasound probe realignment by optical tracking," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, pp. 21–24, IEEE, 2013.

[44] V. Chan and A. Perlas, "Basics of ultrasound imaging," in *Atlas of ultrasound-guided procedures in interventional pain management*, pp. 13–19, Springer, 2011.

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

[46] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, "Machine learning for medical ultrasound: status, methods, and future opportunities," *Abdominal radiology*, vol. 43, no. 4, pp. 786–799, 2018.

[47] B.-S. Lin, J.-L. Chen, Y.-H. Tu, Y.-X. Shih, Y.-C. Lin, W.-L. Chi, and Y.-C. Wu, "Using deep learning in ultrasound imaging of bicipital peritendinous effusion to grade inflammation severity," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 1037–1045, 2020.

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

[49] Z. Wang, Q. Yang, H. Liu, L. Mao, H. Zhu, and X. Gao, "Arb u-net: An improved neural network for suprapatellar bursa effusion ultrasound image segmentation," in *International Conference on Artificial Neural Networks*, pp. 14–23, Springer, 2022.

[50] P. Tyrrell, V. Blanchette, M. Mendez, D. Paniukov, B. Brand, M. Zak, and J. Roth, "Detection of joint effusions in pediatric patients with hemophilia using artificial intelligence-assisted ultrasound scanning; early insights from the development of a self-management tool," *Res Pract Thromb Haemost*, vol. 5, 2021.

[51] D. Ai, C. Cui, Y. Tang, Y. Wang, N. Zhang, C. Zhang, Y. Zhen, G. Li, K. Huang, G. Liu, *et al.*, "Machine learning model for predicting physical activity related bleeding risk in chinese boys with haemophilia a," *Thrombosis Research*, vol. 232, pp. 43–53, 2023.

[52] 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.

[53] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, "Liver fibrosis classification based on transfer learning and fcnet for ultrasound images," *Ieee Access*, vol. 5, pp. 5804–5810, 2017.

[54] H. Tanaka, S.-W. Chiu, T. Watanabe, S. Kaoku, and T. Yamaguchi, "Computer-aided diagnosis system for breast ultrasound images using deep learning," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 235013, 2019.

[55] A. S. Becker, M. Mueller, E. Stoffel, M. Marcon, S. Ghafoor, and A. Boss, "Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study," *The British journal of radiology*, vol. 91, no. xxxx, p. 20170576, 2018.

[56] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning," *Ultrasound in medicine & biology*, vol. 46, no. 5, pp. 1119–1132, 2020.

[57] U. R. Acharya, O. Faust, F. Molinari, S. V. Sree, S. P. Junnarkar, and V. Sudarshan, "Ultrasound-based tissue characterization and classification of fatty liver disease: A screening and diagnostic paradigm," *Knowledge-Based Systems*, vol. 75, pp. 66–77, 2015.

[58] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 919–923, IEEE, 2017.

[59] J. Song, Y. J. Chai, H. Masuoka, S.-W. Park, S.-j. Kim, J. Y. Choi, H.-J. Kong, K. E. Lee, J. Lee, N. Kwak, *et al.*, "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," *Medicine*, vol. 98, no. 15, 2019.

[60] Z. Akkus, J. Cai, A. Boonrod, A. Zeinoddini, A. D. Weston, K. A. Philbrick, and B. J. Erickson, "A survey of deep-learning applications in ultrasound: Artificial intelligence-powered ultrasound for improving clinical workflow," *Journal of the American College of Radiology*, vol. 16, no. 9, pp. 1318–1328, 2019.

[61] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep feature learning for medical image analysis with convolutional autoencoder neural network," *IEEE transactions on big data*, vol. 7, no. 4, pp. 750–758, 2021.

[62] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia computer science*, vol. 132, pp. 377–384, 2018.

[63] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

[64] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of digital imaging*, vol. 30, no. 2, pp. 234–243, 2017.

[65] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," *International journal of advanced computer science & applications*, vol. 10, no. 5, 2019.

[66] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, "Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation," in *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 34–43, Springer, 2023.

[67] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

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

[69] M. Shaha and M. Pawar, "Transfer learning for image classification," in *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pp. 656–660, IEEE, 2018.

[70] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[71] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[72] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, pp. 1–40, 2016.

[73] X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, and Y.-D. Zhang, "Transfer learning for medical images analyses: A survey," *Neurocomputing*, vol. 489, pp. 230–254, 2022.

[74] J. Amin, M. Sharif, M. Yasmin, T. Saba, M. A. Anjum, and S. L. Fernandes, "A new approach for brain tumor segmentation and classification based on score level fusion using transfer learning," *Journal of medical systems*, vol. 43, pp. 1–16, 2019.

[75] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "Covid-19 detection through transfer learning using multimodal imaging data," *Ieee Access*, vol. 8, pp. 149808–149824, 2020.

[76] N. Dhungel, G. Carneiro, and A. P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," *Medical image analysis*, vol. 37, pp. 114–128, 2017.

[77] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020.

[78] T. Clark, J. Zhang, S. Baig, A. Wong, M. A. Haider, and F. Khalvati, "Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted mri using convolutional neural networks," *Journal of Medical Imaging*, vol. 4, no. 4, pp. 041307–041307, 2017.

[79] G. Ayana, K. Dese, and S.-w. Choe, "Transfer learning in breast cancer diagnoses via ultrasound imaging," *Cancers*, vol. 13, no. 4, p. 738, 2021.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common

objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[82] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[83] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*, pp. 1–8, IEEE, 2019.

[84] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1480–1484, IEEE, 2019.

[85] C. Wang, J. Muhammad, Y. Wang, Z. He, and Z. Sun, "Towards complete and accurate iris segmentation using deep multi-task attention network for non-cooperative iris recognition," *IEEE Transactions on information forensics and security*, vol. 15, pp. 2944–2959, 2020.

[86] W. Chen, M. Liu, C. Zhao, X. Li, and Y. Wang, "Mtd-yolo: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection," *Computers and Electronics in Agriculture*, vol. 216, p. 108533, 2024.

[87] Y.-L. Khor, Y. J. Wong, M.-L. Tham, Y. C. Chang, B.-H. Kwan, and K.-C. Khor, "Multi-task yolo for vehicle colour recognition and automatic license plate recognition," in *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 1–7, IEEE, 2024.

[88] K. Yan, Y. Tang, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "Mulan: multitask universal lesion analysis

network for joint lesion detection, tagging, and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 194–202, Springer, 2019.

[89] F. Gao, H. Yoon, T. Wu, and X. Chu, "A feature transfer enabled multi-task deep learning model on medical imaging," *Expert Systems with Applications*, vol. 143, p. 112957, 2020.

[90] M. V. Sainz de Cea, K. Diedrich, R. Bakalo, L. Ness, and D. Richmond, "Multi-task learning for detection and classification of cancer in screening mammography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 241–250, Springer, 2020.

[91] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[92] T.-L.-T. Le, N. Thome, S. Bernard, V. Bismuth, and F. Patoureaux, "Multitask classification and segmentation for cancer diagnosis in mammography," *arXiv preprint arXiv:1909.05397*, 2019.

[93] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, "Multi-task learning for thyroid nodule segmentation with thyroid region prior," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 257–261, IEEE, 2021.

[94] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, and B. Wei, "Sha-mtl: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 10, pp. 1719–1725, 2021.

[95] M. E. Tschuchnig and M. Gadermayr, "Anomaly detection in medical imaging-a mini review," in *International Data Science Conference*, Springer, 2022.

[96] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[97] T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo, "Fraud detection: A systematic literature review of graph-based anomaly detection approaches," *Decision Support Systems*, vol. 133, p. 113303, 2020.

[98] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[99] F. Ye, H. Zheng, C. Huang, and Y. Zhang, "Deep unsupervised image anomaly detection: An information theoretic framework," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1609–1613, IEEE, 2021.

[100] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences. arxiv 2020," *arXiv preprint arXiv:2005.02357*, 2005.

[101] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by cnn-based self-similarity," *Sensors*, vol. 18, no. 1, p. 209, 2018.

[102] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[103] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*, Springer, 2019.

[104] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[105] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

[106] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023.

[107] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

[108] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[109] Y. Mao, F.-F. Xue, R. Wang, J. Zhang, W.-S. Zheng, and H. Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pp. 529–538, Springer, 2020.

[110] J. Tan, B. Hou, T. Day, J. Simpson, D. Rueckert, and B. Kainz, "Detecting outliers with poisson image interpolation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 581–591, Springer, 2021.

[111] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction," *BMC bioinformatics*, vol. 22, pp. 1–20, 2021.

[112] W. H. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, *et al.*, "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 705–714, Springer, 2022.

[113] M. Jiang, C. Hou, A. Zheng, X. Hu, S. Han, H. Huang, X. He, P. S. Yu, and Y. Zhao, "Weakly supervised anomaly detection: A survey," *arXiv preprint arXiv:2302.04549*, 2023.

[114] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[115] Y. Zhao, G. Zheng, S. Mukherjee, R. McCann, and A. Awadallah, "Admoe: Anomaly detection with mixture-of-experts from noisy labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 4937–4945, 2023.

[116] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," in *Medical imaging with deep learning*, pp. 365–381, PMLR, 2020.

[117] C. Rother, V. Kolmogorov, and A. Blake, “” grabcut” interactive foreground extraction using iterated graph cuts,” *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[118] J. Dai, K. He, and J. Sun, “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1635–1643, 2015.

[119] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, *et al.*, “Deepcut: Object segmentation from bounding box annotations using convolutional neural networks,” *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 674–683, 2016.

[120] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, “Adbench: Anomaly detection benchmark,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 32142–32159, 2022.

[121] X. Liu, Z. Liu, Y. Zhang, M. Wang, B. Li, and J. Tang, “Weakly-supervised automatic biomarkers detection and classification of retinal optical coherence tomography images,” in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 71–75, IEEE, 2021.

[122] J. Yang, N. Mehta, G. Demirci, X. Hu, M. S. Ramakrishnan, M. Naguib, C. Chen, and C.-L. Tsai, “Anomaly-guided weakly supervised lesion segmentation on retinal oct images,” *Medical Image Analysis*, vol. 94, p. 103139, 2024.

[123] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion models for medical anomaly detection,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 35–45, Springer, 2022.

[124] J. Li, H. Cao, J. Wang, F. Liu, Q. Dou, G. Chen, and P.-A. Heng, "Fast non-markovian diffusion model for weakly supervised anomaly detection in brain mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 579–589, Springer, 2023.

[125] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, 2011.

[126] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization," in *International Conference on Machine Learning*, pp. 7721–7735, PMLR, 2021.

[127] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.

[128] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22405–22418, 2021.

[129] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

[130] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International conference on machine learning*, pp. 6028–6039, PMLR, 2020.

[131] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization un-

der distribution shifts," in *International conference on machine learning*, pp. 9229–9248, PMLR, 2020.

[132] G. A. V. Hakim, D. Osowiechi, M. Noori, M. Cheraghalikhani, A. Bahri, I. Ben Ayed, and C. Desrosiers, "Clust3: Information invariant test-time training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6136–6145, 2023.

[133] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 213–226, Springer, 2010.

[134] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[135] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek, "Evaluating prediction-time batch normalization for robustness under covariate shift," *arXiv preprint arXiv:2006.10963*, 2020.

[136] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[137] A. T. Nguyen, T. Nguyen-Tang, S.-N. Lim, and P. H. Torr, "Tipi: Test time adaptation with transformation invariance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24162–24171, 2023.

[138] Y. Gandelsman, Y. Sun, X. Chen, and A. Efros, "Test-time training with masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29374–29385, 2022.

[139] D. Osowiechi, G. A. V. Hakim, M. Noori, M. Cheraghalikhani, I. Ben Ayed, and C. Desrosiers, "Tttflow: Unsupervised test-time training with normalizing flow," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2126–2134, 2023.

[140] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

[141] Z. Zhang, W. Chen, H. Cheng, Z. Li, S. Li, L. Lin, and G. Li, "Divide and contrast: Source-free domain adaptation via adaptive contrastive learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5137–5149, 2022.

[142] Y. Liu, P. Kothari, B. Van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi, "Ttt++: When does self-supervised test-time training fail or thrive?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21808–21820, 2021.

[143] D. Osowiechi, G. A. V. Hakim, M. Noori, M. Cheraghalikhani, A. Bahri, M. Yazdanpanah, I. Ben Ayed, and C. Desrosiers, "Nc-ttt: A noise constrastive approach for test-time training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6078–6086, 2024.

[144] C. Fang, S. Bai, Q. Chen, Y. Zhou, L. Xia, L. Qin, S. Gong, X. Xie, C. Zhou, D. Tu, *et al.*, "Deep learning for predicting covid-19 malignant progression," *Medical image analysis*, vol. 72, p. 102096, 2021.

[145] W. Ma, C. Chen, S. Zheng, J. Qin, H. Zhang, and Q. Dou, "Test-time adaptation with calibration of medical image classification nets for label distribution shift," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 313–323, Springer, 2022.

[146] M. Bateson, H. Lombaert, and I. Ben Ayed, "Test-time adaptation with shape moments for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 736–745, Springer, 2022.

[147] J. M. J. Valanarasu, P. Guo, V. Vibashan, and V. M. Patel, "On-the-fly test-time adaptation for medical image segmentation," in *Medical Imaging with Deep Learning*, pp. 586–598, PMLR, 2024.

[148] H. Yang, C. Chen, M. Jiang, Q. Liu, J. Cao, P. A. Heng, and Q. Dou, "Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3575–3586, 2022.

[149] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[150] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?," in *International conference on machine learning*, pp. 5389–5400, PMLR, 2019.

[151] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko, "Visda: A synthetic-to-real benchmark for visual domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2021–2026, 2018.

[152] M. G. Campana, M. Colussi, F. Delmastro, S. Mascetti, and E. Pagani, "Mpox close skin images," May 2023.

[153] Tzutalin, "Labelimg." `https://github.com/tzutalin/labelImg`, 2015.

[154] K. Tingelhoff, K. W. Eichhorn, I. Wagner, M. E. Kunkel, A. I. Moral, M. E. Rilk, F. M. Wahl, and F. Bootz, "Analysis of manual segmentation in paranasal ct images," *European archives of oto-rhino-laryngology*, vol. 265, no. 9, pp. 1061–1070, 2008.

[155] World Health Organization (WHO), "2022 mpox (monkeypox) outbreak: Fact sheets." `https://www.who.int/news-room/fact-sheets/detail/monkeypox`. Accessed: 2023-02-08.

[156] S. N. Ali, M. Ahmed, J. Paul, T. Jahan, S. Sani, N. Noor, T. Hasan, *et al.*, "Monkeypox skin lesion detection using deep learning models: A feasibility study," *arXiv preprint arXiv:2207.03342*, 2022.

[157] X. Wu, W. Ni, L. Jie, Y.-K. Lai, S. Cheng, Dongyu, Ming-Ming, and J. Yang, "Joint acne image grading and counting via label distribution learning," in *IEEE International Conference on Computer Vision*, 2019.

[158] L. Muñoz-Saavedra, E. Escobar-Linero, J. Civit-Masot, F. Luna-Perejón, A. Civit, and M. Domínguez-Morales, "Monkeypox diagnostic-aid system with skin images using convolutional neural networks," *Available at SSRN 4186534*, 2024.

[159] E. Callaway *et al.*, "Fast-spreading covid variant can elude immune responses," *Nature*, vol. 589, no. 7843, pp. 500–501, 2021.

[160] C.-C. Lai, C.-K. Hsu, M.-Y. Yen, P.-I. Lee, W.-C. Ko, and P.-R. Hsueh, "Monkeypox: An emerging global threat during the covid-19 pandemic," *Journal of Microbiology, Immunology and Infection*, vol. 55, no. 5, pp. 787–794, 2022.

[161] E. M. Bunge, B. Hoet, L. Chen, F. Lienert, H. Weidenthaler, L. R. Baer, and R. Steffen, "The changing epidemiology of human monkeypox—a potential threat? a systematic review," *PLOS Neglected Tropical Diseases*, vol. 16, pp. 1–20, 02 2022.

[162] P. v. Magnus, E. K. Andersen, K. B. Petersen, and A. Birch-Andersen, “A pox-like disease in cynomolgus monkeys,” *Acta Pathologica Microbiologica Scandinavica*, vol. 46, no. 2, pp. 156–176, 1959.

[163] J. G. Rizk, G. Lippi, B. M. Henry, D. N. Forthal, and Y. Rizk, “Prevention and treatment of monkeypox,” *Drugs*, vol. 82, pp. 957–963, Jun 2022.

[164] C. for Disease Control and P. (CDC), “About mpox.” `https://www.cdc.gov/poxvirus/monkeypox/about`. Accessed: 2023-02-08.

[165] A. Asadzadeh and L. R. Kalankesh, “A scope of mobile health solutions in covid-19 pandemics,” *Informatics in Medicine Unlocked*, vol. 23, p. 100558, 2021.

[166] V. K. Rajendran, P. Bakthavathsalam, P. L. Bergquist, and A. Sunna, “Smartphone technology facilitates point-of-care nucleic acid diagnosis: a beginner’s guide,” *Critical Reviews in Clinical Laboratory Sciences*, vol. 58, no. 2, pp. 77–100, 2021.

[167] Z. Rong, Q. Wang, N. Sun, X. Jia, K. Wang, R. Xiao, and S. Wang, “Smartphone-based fluorescent lateral flow immunoassay platform for highly sensitive point-of-care detection of zika virus nonstructural protein 1,” *Analytica Chimica Acta*, vol. 1055, pp. 140–147, 2019.

[168] P. Brangel, A. Sobarzo, C. Parolo, B. S. Miller, P. D. Howes, S. Gelkop, J. J. Lutwama, J. M. Dye, R. A. McKendry, L. Lobel, and M. M. Stevens, “A serological point-of-care test for the detection of igg antibodies against ebola virus in human survivors,” *ACS Nano*, vol. 12, pp. 63–73, Jan 2018.

[169] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat,

A. Floto, P. Cicuta, and C. Mascolo, "Sounds of covid-19: exploring realistic performance of audio-based digital testing," *npj Digital Medicine*, vol. 5, p. 16, Jan 2022.

[170] M. G. Campana, A. Rovati, F. Delmastro, and E. Pagani, "L3-net deep audio embeddings to improve covid-19 detection from smartphone data," in *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 100–107, 2022.

[171] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[172] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022.

[173] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing* (J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, eds.), (Cham), pp. 563–574, Springer International Publishing, 2019.

[174] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.

[175] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[176] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[177] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[178] 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.

[179] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[180] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

[181] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[182] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[183] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[184] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, “Mnasnet: Platform-aware neural architecture search for mobile,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[185] M. Huh, P. Agrawal, and A. A. Efros, “What makes imagenet good for transfer learning?,” *arXiv preprint arXiv:1608.08614*, 2016.

[186] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[187] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[188] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[189] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” *arXiv preprint arXiv:1510.00149*, 2015.

[190] A. Kwasniewska, M. Szankin, M. Ozga, J. Wolfe, A. Das, A. Zajac, J. Ruminski, and P. Rad, “Deep learning optimization for edge devices: Analysis of training quantization parameters,” in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1, pp. 96–101, 2019.

[191] R. Ibrahim and M. O. Shafiq, "Explainable convolutional neural networks: A taxonomy, review, and future directions," *ACM Comput. Surv.*, vol. 55, feb 2023.

[192] 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.

[193] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[194] P. Bourdon, O. B. Ahmed, T. Urruty, K. Djemal, and C. Fernandez-Maloigne, "Explainable ai for medical imaging: Knowledge matters," in *Multi-faceted Deep Learning*, pp. 267–292, Springer, 2021.

[195] T. B. Fitzpatrick, "The Validity and Practicality of Sun-Reactive Skin Types I Through VI," *Archives of Dermatology*, vol. 124, pp. 869–871, 06 1988.

[196] G. A. Tadesse, C. Cintas, K. R. Varshney, P. Staar, C. Agunwa, S. Speakman, J. Jia, E. E. Bailey, A. Adelekun, J. B. Lipoff, *et al.*, "Skin tone analysis for representation in educational materials (star-ed) using machine learning," *NPJ Digital Medicine*, vol. 6, no. 1, p. 151, 2023.

[197] M. Wilkes, C. Y. Wright, J. L. du Plessis, and A. Reeder, "Fitzpatrick Skin Type, Individual Typology Angle, and Melanin Index in an African Population: Steps Toward Universally Applicable Skin Photosensitivity Assessments," *JAMA Dermatology*, vol. 151, pp. 902–903, 08 2015.

[198] I. UVA, "Uva1-induced skin darkening is associated with molecular changes even in highly pigmented skin individuals," *Journal of Investigative Dermatology*, vol. 137, p. 1184e1187, 2017.

[199] X. Li, G. Zhang, H. H. Huang, Z. Wang, and W. Zheng, "Performance analysis of gpu-based convolutional neural networks," in *2016 45th International Conference on Parallel Processing (ICPP)*, pp. 67–76, 2016.

[200] A. Mohiyuddin, A. Basharat, U. Ghani, S. Abbas, O. B. Naeem, and M. Rizwan, "Breast tumor detection and classification in mammogram images using modified yolov5 network," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[201] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, "ultralytics/yolov5: v6.1 - tensorrt, tensorflow edge tpu and openvino export and inference," feb 2022.

[202] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[203] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, 2020.

[204] K. Jabeen, M. A. Khan, M. Alhaisoni, U. Tariq, Y.-D. Zhang, A. Hamza, A. Mickus, and R. Damaševičius, "Breast cancer classification from ultrasound images using probability-based

optimal deep learning feature fusion," *Sensors*, vol. 22, no. 3, p. 807, 2022.

[205] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv.org*, 2018.

[206] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 12993–13000, 2020.

[207] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary intelligence*, vol. 15, no. 1, pp. 1–22, 2022.

[208] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[209] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*, pp. 3121–3124, IEEE, 2010.

[210] B. Ci and R.-O. Rule, "Confidence intervals," *Lancet*, vol. 1, no. 8531, pp. 494–7, 1987.

[211] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, PMLR, 2013.

[212] E. Bochinski, T. Senst, and T. Sikora, "Hyper-parameter optimization for convolutional neural network committees based on evolutionary algorithms," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3924–3928, IEEE, 2017.

[213] M. Power, G. Fell, and M. Wright, "Principles for high-quality, high-value testing," *BMJ Evidence-Based Medicine*, vol. 18, no. 1, pp. 5–10, 2013.

[214] N. Schenker and J. F. Gentleman, "On judging the significance of differences by examining the overlap between confidence intervals," *The American Statistician*, vol. 55, no. 3, pp. 182–186, 2001.

[215] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[216] R. El Jurdi, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah, "High-level prior-based loss functions for medical image segmentation: A survey," *Computer Vision and Image Understanding*, vol. 210, p. 103248, 2021.

[217] V. Kulharia, S. Chandra, A. Agrawal, P. Torr, and A. Tyagi, "Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation," in *European Conference on Computer Vision*, pp. 290–308, Springer, 2020.

[218] G. Feng, L. Zhang, Z. Hu, and H. Lu, "Learning from box annotations for referring image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[219] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[220] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE transactions on image processing*, vol. 23, no. 2, p. 684–695, 2014.

[221] F. Peyvandi, I. Garagiola, and E. Biguzzi, “Advances in the treatment of bleeding disorders,” *Journal of Thrombosis and Haemostasis*, vol. 14, no. 11, pp. 2095–2106, 2016.

[222] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015.

[223] B. Zhang, P. V. Sander, and A. Bermak, “Gradient magnitude similarity deviation on multiple scales for color image quality assessment,” in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1253–1257, IEEE, 2017.

[224] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[225] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, “Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[226] J. Pirnay and K. Chai, “Inpainting transformer for anomaly detection,” in *International Conference on Image Analysis and Processing*, pp. 394–406, Springer, 2022.

[227] T. Liu, B. Li, X. Du, B. Jiang, L. Geng, F. Wang, and Z. Zhao, “Fair: Frequency-aware image restoration for industrial visual anomaly detection,” *arXiv preprint arXiv:2309.07068*, 2023.

[228] Z. Zong, G. Song, and Y. Liu, “Detrs with collaborative hybrid assignments training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

[229] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, “Embracing imperfect datasets: A review of deep

learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[230] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *Computers in Biology and Medicine*, p. 107912, 2023.

[231] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[232] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[233] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[234] J. Guo, L. Jia, W. Zhang, H. Li, *et al.*, "Recontrast: Domain-specific anomaly detection via contrastive reconstruction," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[235] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

[236] P. Chrabaszcz, I. Loshchilov, and F. Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.

[237] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, and C. Finn, "Surgical fine-tuning improves adaptation to distribution shifts," *arXiv preprint arXiv:2210.11466*, 2022.

[238] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2427–2440, 2021.

[239] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[240] N. Y. Philip, J. J. Rodrigues, H. Wang, S. J. Fong, and J. Chen, "Internet of things for in-home health monitoring systems: Current advances, challenges and future directions," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 300–310, 2021.

[241] R. Gualtierotti, S. Arcudi, A. Ciavarella, M. Colussi, S. Mascetti, C. Bettini, and F. Peyvandi, "A computer-aided diagnosis tool for the detection of hemarthrosis by remote joint ultrasound in patients with hemophilia," *Blood*, vol. 140, no. Supplement 1, pp. 464–465, 2022.

[242] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[243] L. A. Valentino, "Considerations in individualizing prophylaxis in patients with haemophilia a," *Haemophilia*, vol. 20, no. 5, pp. 607–615, 2014.

[244] J. O'Hara, D. Hughes, C. Camp, T. Burke, L. Carroll, and D.-A. G. Diego, "The cost of severe haemophilia in europe: the

chess study," *Orphanet journal of rare diseases*, vol. 12, pp. 1–8, 2017.

[245] N. Painchaud, N. Duchateau, O. Bernard, and P.-M. Jodoin, "Echocardiography segmentation with enforced temporal consistency," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2867–2878, 2022.

[246] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2024. Open source software available from https://github.com/heartexlabs/label-studio.