Intelligent Systems with Applications 17 (2023) 200183

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Ultrasound detection of subquadricipital recess distension

Marco Colussi [a,*], Gabriele Civitarese [a], Dragan Ahmetovic [a], Claudio Bettini [a], Roberta Gualtierotti [b,c], Flora Peyvandi [b,c], Sergio Mascetti [a]

[a] Università degli Studi di Milano, Department of Computer Science, Via Celoria, 18, 20133, Milan, Italy
[b] Università degli Studi di Milano, Department of Pathophysiology and Transplantation, Via Pace, 9, 20122, Milan, Italy
[c] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, Via Pace, 9, 20122, Milan, Italy

ARTICLE INFO

Keywords:
Multi-task learning
Clinical decision support
Ecography
Hemarthrosis

ABSTRACT

Joint bleeding is a common occurrence for people with hemophilia and, if untreated, can result in hemophilic arthropathy. Ultrasound imaging has recently emerged as an effective tool to diagnose the distension of a joint recess caused by joint bleeding. However, no computer-aided diagnosis tool exists to support the practitioner in the diagnosis process. This paper addresses the problem of automatically detecting the subquadricipital recess of the knee and assessing whether it is distended in ultrasound images collected from patients with hemophilia. After framing the problem, we propose two different approaches: the first one adapts a one-stage object detection algorithm for the task, while the second one is a multi-task approach with a classification and a detection branch. The experimental evaluation, conducted with 483 annotated images, shows that the solution based on object detection alone has a balanced accuracy score of 0.74 with a mean IoU value of 0.66, while the multi-task approach has a higher balanced accuracy value (0.78) at the cost of a slightly lower mean IoU value.

1. Introduction

Hemophilia is a hereditary blood coagulation disorder that results in an increased risk of bleeding, due to trauma or spontaneously, which worsens with the severity of the disease. Bleedings can frequently occur also inside joints (mostly ankles, knees and elbows) and muscles, which together account for around 80% of the bleeding events in patients with Hemophilia (Roosendaal & Lafeber, 2003, Srivastava et al., 2020). Joint bleeding causes the *distension* of the affected joint recess and, if recurrent, can result in synovial hyperplasia, osteochondral damage, and hemophilic arthropathy (Hilgartner, 2002). Thus, it is essential to promptly recognize joint recess distension.

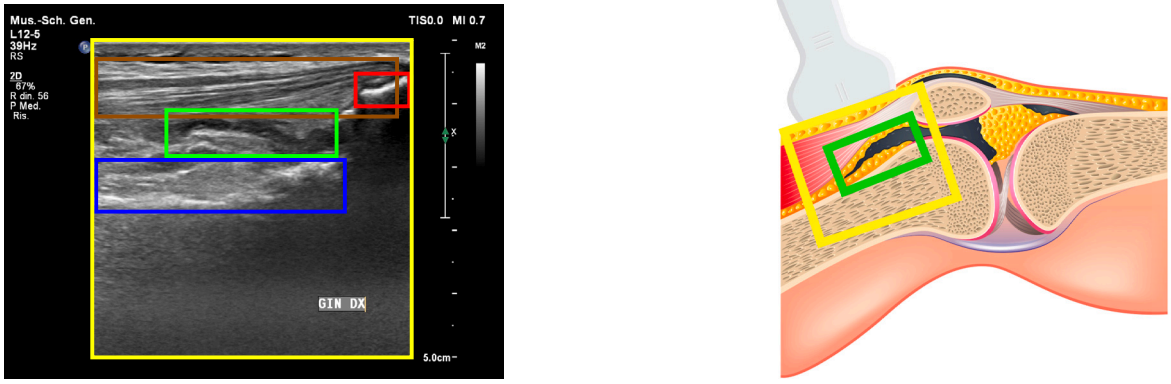
Physical examination may not be sufficient to diagnose joint recess distension, since in the early stage it can be asymptomatic (Plut et al., 2019). Magnetic Resonance Imagining (MRI) is generally considered the gold standard tool for precise evaluation of joints but it is not practical for regular follow-up of patients with hemophilia due to the high costs, limited availability and long examination times (Plut et al., 2019). An alternative solution is ultrasound (US) imaging (Wells, 2006) that, contrary to MRI, has a low cost, short examining time and it is widely accessible (Joshua et al., 2007). *Hemophilia Early Arthropathy Detection with UltraSound* (HEAD-US) is a standardized protocol designed to guide the practitioner in acquiring relevant US images and interpreting them for the diagnosis of joint recess distension in the 6 most commonly affected joints (Martinoli et al., 2013).

Computer aided diagnosis (CAD) systems can improve detection accuracy (Chan et al., 1990) and reduce the image reading time required by the practitioners (Doi, 2005). The potential effectiveness of US-based CAD systems to support the diagnosis of joint distension in people with hemophilia is suggested by recent studies that focus in identifying joint health related to injuries (Long et al., 2020).

In this work, we formulate the research problem of supporting the physicians in diagnosing joint recess distension in patients with hemophilia using a CAD system. The problem consists of detecting the joint recess inside 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 HEAD-US protocol for this joint (Martinoli et al., 2013). One prior work addresses the problem of detecting SQR distension in pediatric patients with hemophilia (Tyrrell et al., 2021), but specific details about the methodology and the evaluation are not reported.

* Corresponding author.
E-mail addresses: marco.colussi@unimi.it (M. Colussi), gabriele.civitarese@unimi.it (G. Civitarese), dragan.ahmetovic@unimi.it (D. Ahmetovic), claudio.bettini@unimi.it (C. Bettini), roberta.gualtierotti@unimi.it (R. Gualtierotti), flora.peyvandi@unimi.it (F. Peyvandi), sergio.mascetti@unimi.it (S. Mascetti).

https://doi.org/10.1016/j.iswa.2023.200183
Received 26 July 2022; Received in revised form 6 December 2022; Accepted 14 January 2023
Available online 20 January 2023



(a) Example of SQR longitudinal scan

(b) Probe positioning

Fig. 1. Image acquisition.

Besides formulating the research problem, we also propose two approaches to address it. 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 a new dataset of images that we collected and annotated due to the lack of other publicly available datasets. In our dataset, we collected more than 450 images from 208 adult subjects with hemophilia, which is consistent with other studies in the literature (Tyrrell et al., 2021, Wang et al., 2022, Long et al., 2020). In the experiments we compared the two proposed solutions among themselves and with two baselines, one *Classification baseline* and one *Detection baseline*. 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. For what concerns the detection accuracy, the *Detection approach* has a slightly better performance than the *Multi-task approach*, and it remains in line with the *Detection baseline*.

To sum up, the novel contributions of this paper are the following:

- We formulate the problem of detecting and classifying SQR distensions from US images.
- We propose two solutions to tackle this problem.
- We evaluate and compare the proposed solutions on a dataset collected from 208 patients.

2. Problem formulation

In this research, we address the problem of the automated detection of the SQR recess and its classification as *Distended* or *Non-distended*.

2.1. Ultrasound images

Ultrasound (US) (Chan & Perlas, 2011) is a very popular medical imaging technique. It is portable, safe and affordable and therefore commonly used in healthcare (Brattain et al., 2018). However, some limitations of this technique are the high dependence on the operator expertise level and possible noisiness of the acquired images (Plut et al., 2019).

US imaging uses a sound wave signal at high frequencies. The reflections of the signal are then measured to represent the image. This technique can produce images with a high spatial resolution of the internal structures of the body, like tendons, bones, blood, and muscles (Wells, 2006). The images are represented in grayscale, where each pixel value describes the density of the material the signal encounters. Light areas represent echogenic tissues (*i.e.*, that reflect sound waves) like bones, while dark areas represent anechogenic (*i.e.*, that do not reflect sound) structures such as liquids. Another effect to take into account is that echogenic tissues, such as bone, shield the signal that is unable to travel through them, thus making it impossible to detect anything below them. An example is shown in Fig. 1: the patella is clearly distinguishable in light color (see the red box) while the area below it is almost completely black.

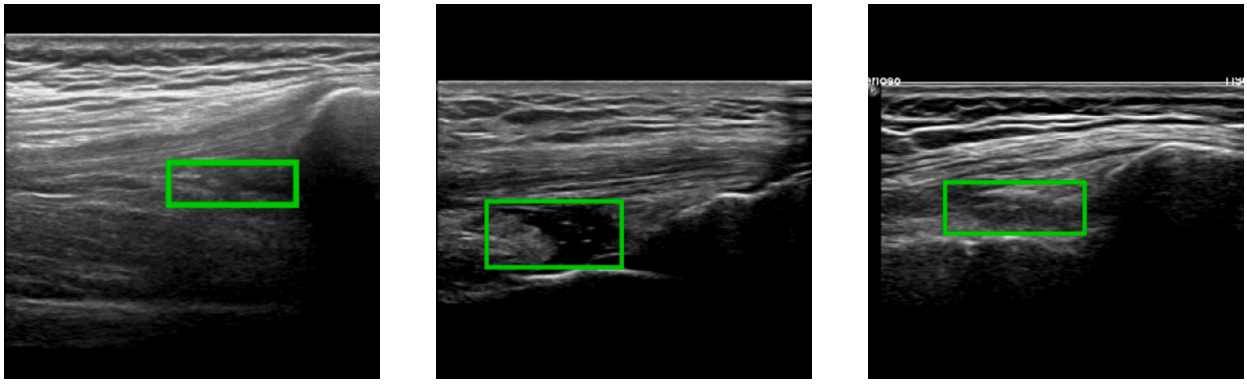
2.2. SQR longitudinal scan

We focus on one of the three scans of the knee joint specified in HEAD-US protocol for the collection and diagnosis of joint recess distension in patients with hemophilia (Martinoli et al., 2013): the SQR longitudinal scan. This scan is used to assess SQR distension and contains different characterizing elements (see Fig. 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.

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 not, as explained below.

Fig. 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 that is captured by the US image shown in Fig. 1a, while the green box is the SQR. To correctly acquire this type of image, the knee has to



(a) Non-distended SQR (b) Distended SQR (c) Borderline Non-distended SQR

Fig. 2. Examples of longitudinal SQR scans.

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.

A number of ultrasound probe parameters need to be specified in order to properly acquire SQR longitudinal scans. Some of these parameters need to be personalized for each patient (like gain, focus and dynamic range), while the value for other parameters can be predetermined, like frequency and depth that in our study were set to $12\ MHz$ and $40-50\ mm$, respectively.

2.3. Detection and classification of SQR distension

A joint recess can be distended due to three main reasons: if it is filled with synovial liquid, if it is filled with blood (a condition known as *hemarthrosis*), and if its membrane is thicker due to an inflammation known as *synovitis*. The approximate recess position can be inferred from the location of the three characterizing elements described in Section 2.2 (*i.e.*, patella, femur, and tendons): the recess is positioned below the tendons, above the right-most end of the femur and on the patella bottom-left. To determine the exact position of the recess, the practitioner observes the anechogenic area that is present in the region. The recess appears as a dark area surrounded by a lighter membrane.

To determine whether the SQR is distended, we rely on the assessment of the US image by a practitioner. The practitioner observes the recess and qualitatively establishes whether it is swollen, a sign that it is filled with liquid or that its membrane is thickened. Instead, a non-distended recess should appear 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 (Long et al., 2020). Indeed, while two alternative approaches are possible, they are impractical: MRI is expensive and time consuming (Plut et al., 2019) while the aspiration of the liquid through puncture (arthrocentesis) is invasive, particularly in patients with bleeding disorders such as hemophilia (Peyvandi et al., 2016).

Fig. 2 shows three examples of the longitudinal SQR scan. In Fig. 2a the SQR is the dark area shown in the green box. In this case the SQR is thin, hence it is not distended. Vice versa, in Fig. 2b the SQR is much thicker, indicating that it is distended. While Fig. 2a and 2b show two characteristic examples with stark differences, there are borderline cases where the SQR appears slightly enlarged but it is not distended (see Fig. 2c) or it is very slightly distended.

An interview, conducted with physicians from the Angelo Bianchi Bonomi Hemophilia and Thrombosis Center (two of which are also authors of this paper), 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 (Gualtierotti et al., 2021, Plut et al., 2019). 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 appearsf 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*.

2.4. Problem modeling

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, 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 (Mohiyuddin et al., 2022). 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.

3. Methodology

We propose two solutions for the problem defined in Section 2. The first solution, which we name *Detection approach*, is described in Section 3.1. 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 3.2), is a multi-task network with a branch that solves the detection problem and another one that solves the classification problem.

3.1. Detection approach

Fig. 3 depicts the network architecture of the *Detection approach*. Each input US image is processed by the YoloV5 (Jocher et al., 2022) object detector that returns a set of candidate SQRs, each characterized by a confidence value, a bounding box and the label (*Distended* or

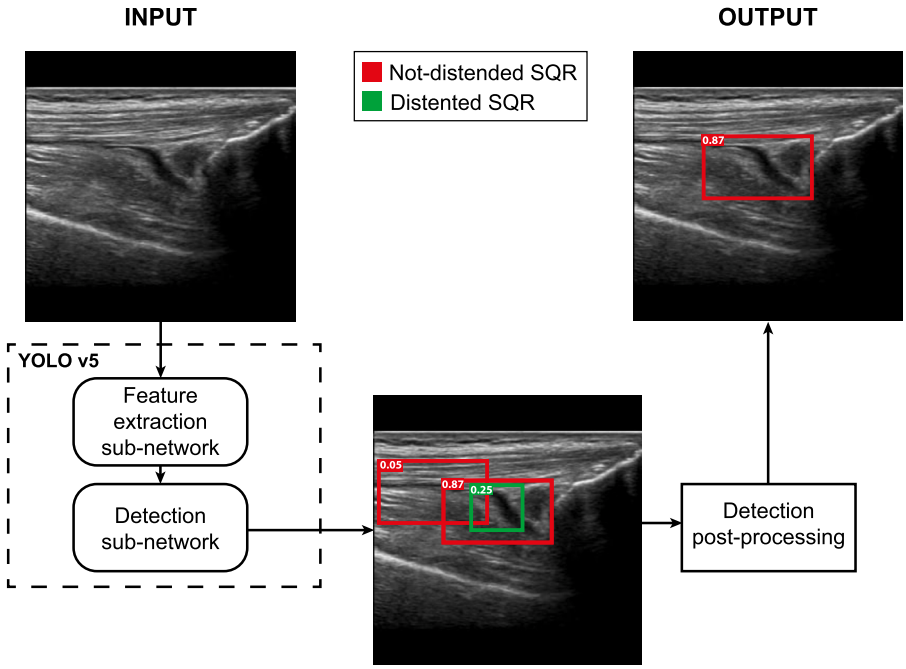


Fig. 3. Overall architecture of the *Detection approach*.

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 (Cheng & Malhi, 2017) 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 (Lin et al., 2014). Finally, the network is fine-tuned on the actual dataset containing the labeled US images.

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 (Bochkovskiy et al., 2020), 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 C.-Y. Wang et al. (2020) and that was shown to be particularly effective for object detection (Bochkovskiy et al., 2020) and ultrasound image classification (Jabeen et al., 2022).

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 (Redmon & Farhadi, 2018) 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 (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) (Zheng et al., 2020), 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 5.4).

$$L = \alpha L_{box} + \beta L_{obj} + \gamma L_c \tag{1}$$

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.

3.2. 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 (see Fig. 4).

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

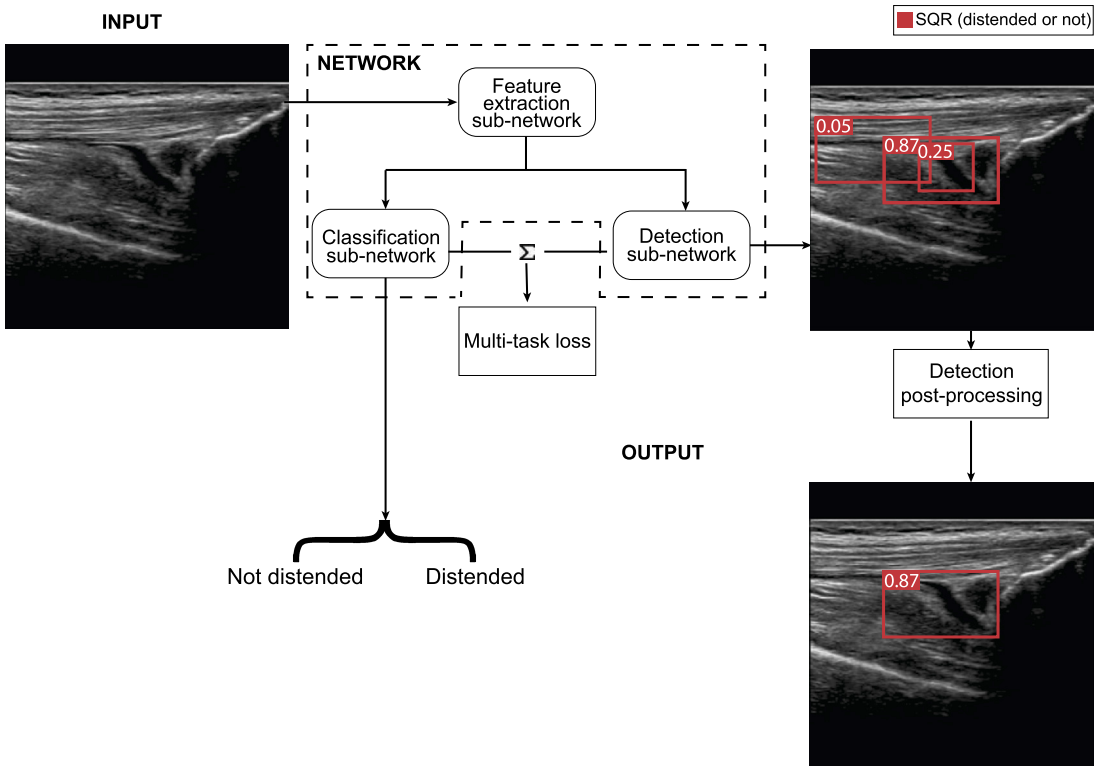


Fig. 4. Overall architecture of the *Multi-task approach*.

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.

3.2.1. Classification sub-network

Fig. 5 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.

3.2.2. 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

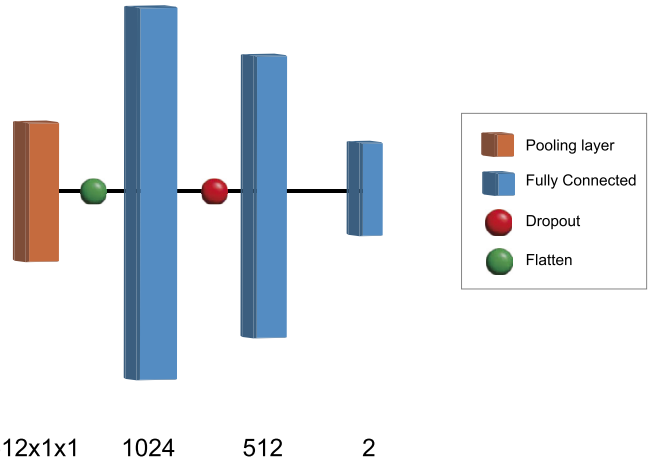


Fig. 5. Classification sub-network architecture.

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 (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 (2). These weights are hyper-parameters that need to be empirically tuned (see Section 5.4).

$$L = \alpha L_{box} + \beta L_{obj} + \delta L_{cls} \tag{2}$$

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.

3.2.3. 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. Dataset

Despite the fact that there are prior works that analyze US images of the relevant area (SQR scan of the knee) (Tyrrell et al., 2021, Wang et al., 2022, Long et al., 2020), 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 adult patients 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 polyclinic of Milan, a medical institution specialized in hemophilia. Images were annotated by expert physicians specifically trained on the diagnosis of the distention of the SQR 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 (Martinoli et al., 2013) 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 data extraction from the ultrasound device; 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), hence having up to four images each.

4.1. Data acquisition and annotation

Images were acquired using the *Philips Affiniti 50* US device[1] by a single specialized practitioner during routine visits of hemophilic patients. When collecting the images, the probe was positioned as shown in Fig. 1b and the knee was flexed by 30°. Each image has a resolution of 1024×780 and, as shown in Fig. 1a, it contains acquisition parameters (saved as text in the image) and the actual US scan (*i.e.*, the yellow rectangle in Fig. 1a), the size of which can vary.

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

The annotation procedure is organized into three phases. The first 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, like 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 (Tzutalin, 2015), the practitioner identifies the SQR position with the approach presented in Section 2.3 and draws the bounding box (a rectangle with edges parallel to the axes).

The third phase is class labeling: using the approach presented in Section 2.3, 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*.

4.2. Pre-processing

We pre-process the collected images to extract the actual US image (e.g., the yellow box in Fig. 1a). Indeed, as previously observed (Lin et al., 2020, Long et al., 2020) 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. (2008), 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. Fig. 6 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 crop 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.

5. 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.

5.1. 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* (Redmon & Farhadi, 2018) 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 (Sarvamangala & Kulkarni, 2022).

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

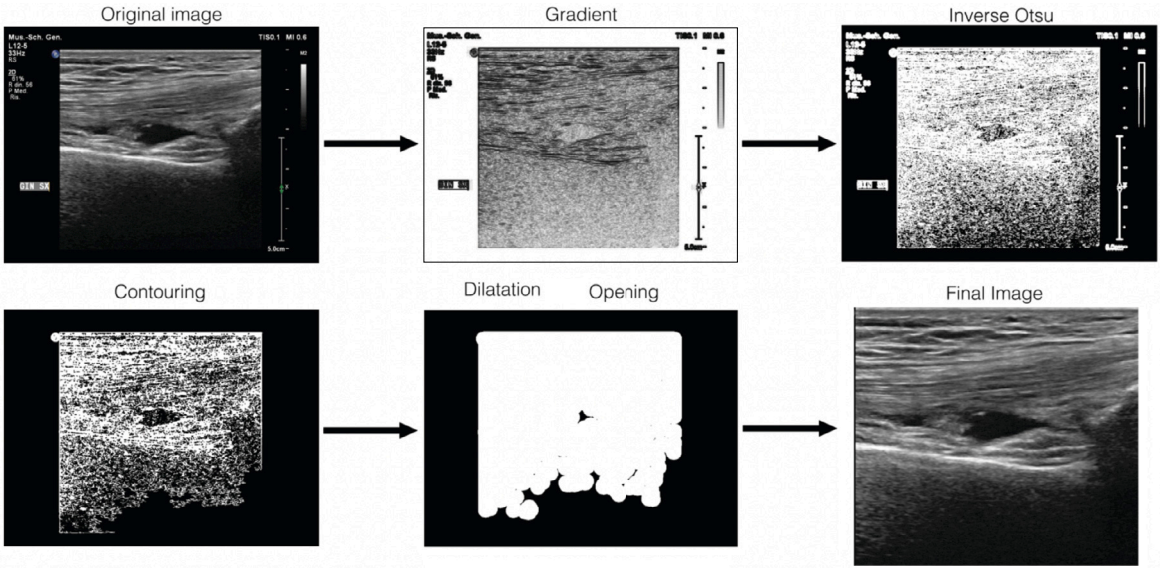


Fig. 6. Intermediate steps of frame extraction procedure.

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 Sarvamangala and Kulkarni (2022) 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 5.3. For the same reason, all four techniques are evaluated using the same cross-validations splits.

5.2. 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 (Everingham et al., 2010). 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 (Brodersen et al., 2010). Balanced accuracy is defined as $\frac{sens+spec}{2}$

Table 1
Example data distribution in Fold 0 of the 5-fold cross-validation.

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

- 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 (Ci & Rule, 1987).

5.3. 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 1. Each training fold was further split: 80% as training set and 20% as validation set. During training we used SGD with momentum (Sutskever et al., 2013) as optimizer.

5.4. Hyper-parameters selection

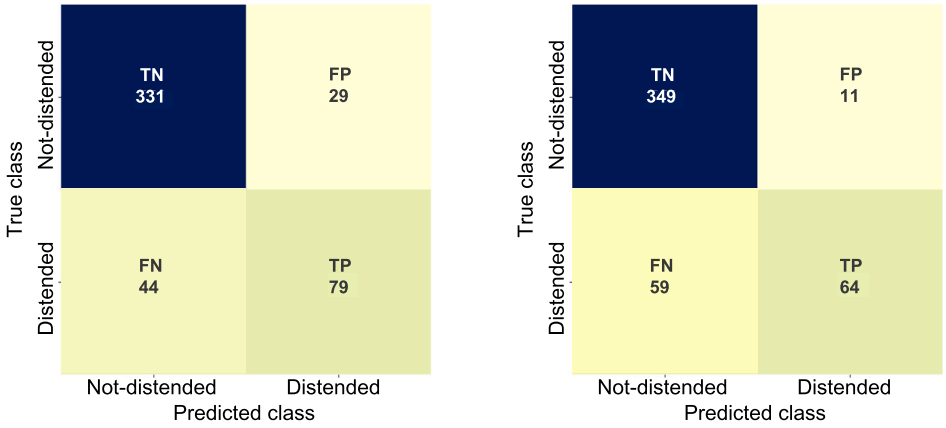
In order to properly tune the many hyper-parameters of our network, we adopt an evolutionary approach (Bochinski et al., 2017). Given a fitness function, an evolutionary algorithm evaluates the best fitting set of hyper-parameters thanks to *mutation* and *cross-over* operations. For the sake of this 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 3.1 and 3.2.3, respectively.

Table 2
Selected hyper-parameters.

	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 3
Evaluation results (reported as mean among the folds ± standard deviation).

	Balanced accuracy	Specificity	Sensitivity	IoU
Classification baseline	0.73 ± 0.03	0.85 ±0.09	0.61 ± 0.13	–
Detection baseline	–	–	–	**0.66 ± 0.02**
Detection Approach	0.74 ± 0.07	**0.97 ±0.03**	0.52 ± 0.12	**0.66 ± 0.01***
Multi-task Approach	**0.78 ± 0.05**	0.92 ± 0.04	**0.64 ± 0.09**	0.63 ± 0.02



(a) Multi-task approach (b) Detection approach

Fig. 7. Confusion matrices.

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 2.

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

5.5. Results

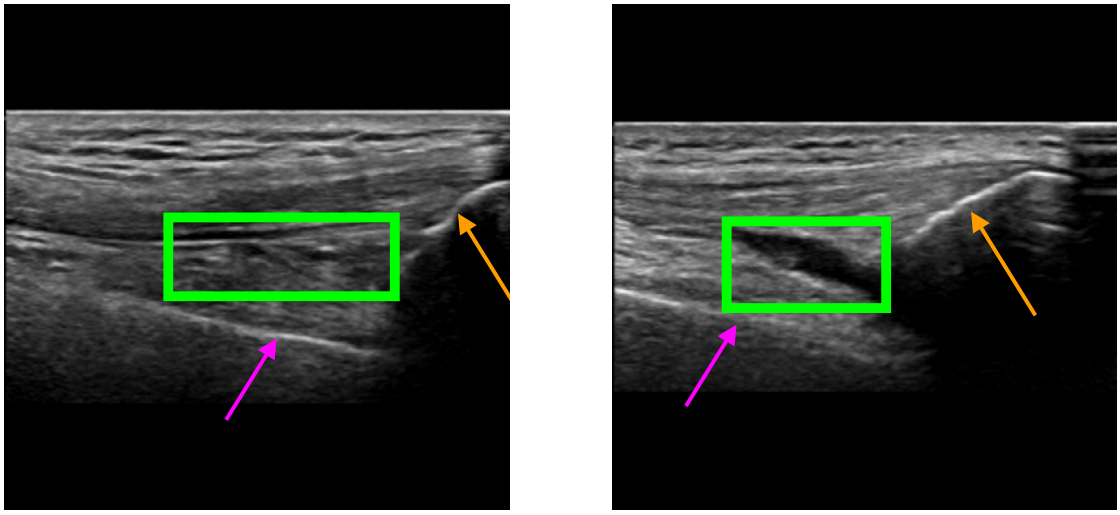
Table 3 shows the performance of the two baselines and of the two proposed solutions. Note that, in order to fairly compare the *Detection approach* with the *Detection baseline* and the *Multi-task approach*, the average IoU for the *Detection approach* (marked with *) is computed ignoring the predicted 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.

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*.

As discussed in Section 5.7, 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 (Power et al., 2013, Everingham et al., 2010). 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 (Schenker & Gentleman, 2001). 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 3.2.2). Indeed, considering the confusion matrices in Fig. 7, 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, as we motivate in Section 5.7.

5.6. Examples

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



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

Fig. 8. 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.

Fig. 8 shows two US images that have been correctly classified by both approaches and that are relatively easy to classify by medical experts. Fig. 8a shows an US image where the femur, the patella and the SQR are clearly visible, and the SQR is thin (*i.e.*, not distended). On the other hand, Fig. 8b shows an example of a *Distended* SQR. In this case, the SQR is clearly thick and hence distended.

Fig. 9 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 Fig. 9c) or when the SQR is borderline between *Distended* and *Non-distended* (as in Fig. 9d). Fig. 9a is correctly classified by both approaches as *Non-distended*. Fig. 9b is correctly classified by the *Multi-task approach* but not by the *Detection approach*. Vice versa, Fig. 9c is correctly classified by the *Detection approach* and not by the *Multi-task approach*. Finally, both solutions wrongly classify Fig. 9d.

Considering the detection problem, Fig. 10 shows US images where the two approaches detected the SQR with the lowest and the highest IoU. In Fig. 10a, 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 portion of the actual SQR (IoU = 0.05). Instead, in the example shown in Fig. 10b 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).

Fig. 10c shows the US image for which the *Detection approach* provided the lowest IoU value. The problem is similar to that of Fig. 10a: 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).

Fig. 10d 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.

5.7. Discussion

The experimental evaluation shows that the *Multi-task approach* results in a better balanced accuracy compared to the *Detection approach*. This is particularly relevant for two reasons. First, the balanced accuracy confidence interval of the *Multi-task approach* is completely above the threshold of 0.75 that is reported to be a requirement for a medical test to be "useful" (Power et al., 2013). 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 (+12%) sensitivity value with respect to the *Detection approach* at the cost of a lower (−5%) specificity value. This is particularly important because, in the considered domain, sensitivity should be privileged over specificity. Indeed, false negatives (captured by sensitivity) have 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 raise the practitioner's attention when not needed and, in the worse scenario, can lead to over-treatment (*e.g.*, provide factor VIII when not actually 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 (Hilgartner, 2002).

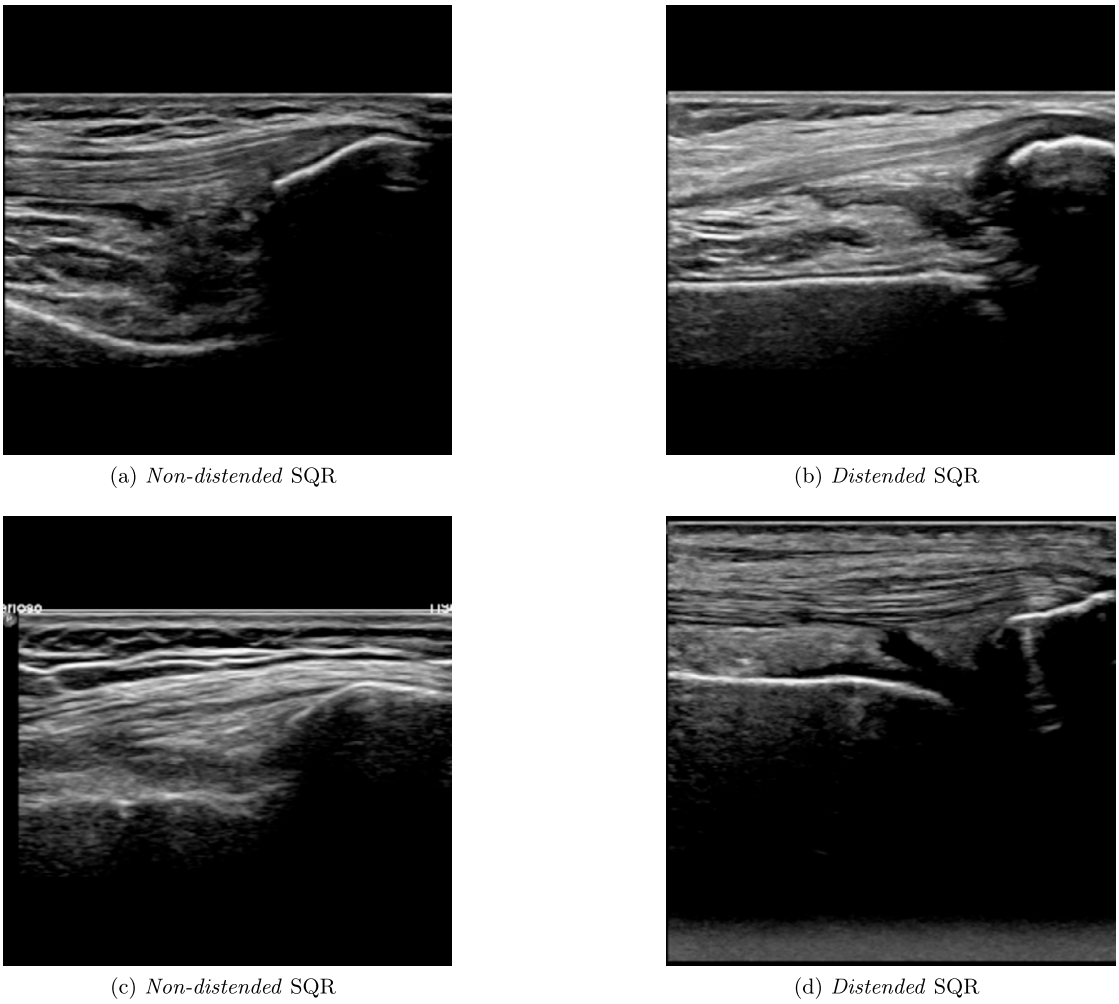
Considering the detection performance, we can observe that when the IoU is above 0.5, which is a common threshold to define for a "correct" detection, the SQR is intuitively correctly detected and hence can support the practitioner during the examination. For example in Fig. 10d the red box has an IoU of 0.55 and indeed it correctly detects the right area, although the bounding rectangle is slightly shorter and larger that the ground truth. With both proposed solutions the IoU is above 0.5 in more than 82% of the cases (85% with *Detection approach* and 82% with *Multi-task approach*). 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 few cases in which the techniques detect the wrong target, as in the examples of Figs. 10a and 10c.

6. Related work

In this section we first report the related work in the broad field of US-based CAD systems. Then, in Section 6.2, we report the existing literature about the classification and detection of joint recess distention in US images and we compare existing works with our solutions.

6.1. US-based CAD systems

Machine Learning (ML) techniques using medical imaging data have been investigated to support physicians in diagnosing various conditions (Fujita, 2020). In particular, Ultrasound (US) (Chan & Perlas, 2011) is a very popular medical imaging technique, often used also as a



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

(c) *Non-distended* SQR (d) *Distended* SQR

Fig. 9. Examples of images that are intuitively hard to classify.

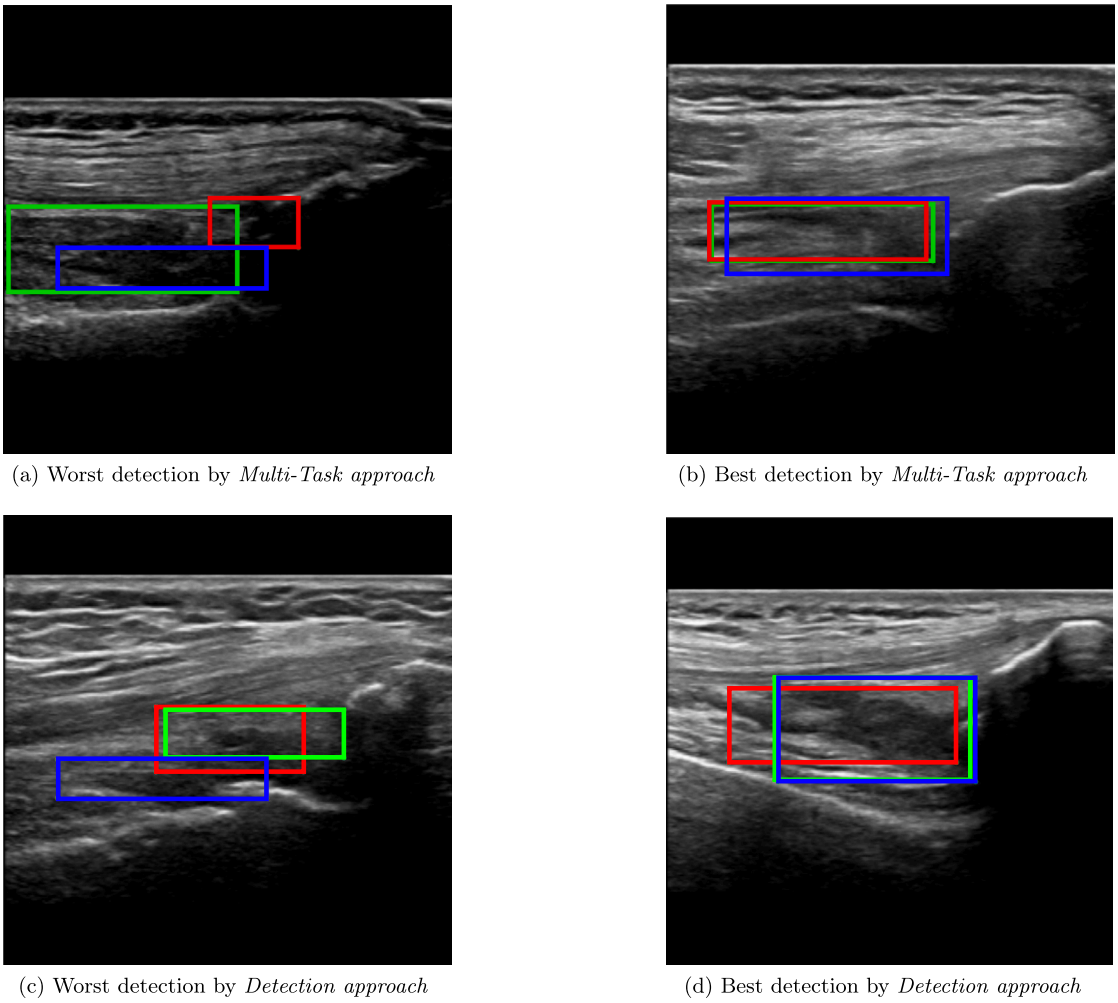
data source for Computer-Aided Diagnosis (CAD) systems (Huang et al., 2018, Brattain et al., 2018). Indeed, despite its high dependence on the operator expertise level and possible noisiness of the acquired images (Plut et al., 2019), US imaging is easily accessible, safe and affordable and therefore commonly used in healthcare (Brattain et al., 2018).

In this problem domain, Convolutional Neural Networks (CNNs) are the most frequently used ML architectures, due to their ability to extract discriminative features from image data (Chen et al., 2021, Simonyan & Zisserman, 2014, Akkus et al., 2019, Sharma et al., 2018). However, the development of such systems is often limited by the scarcity of available labeled data for the training of the ML models. To mitigate this issue, in the literature, transfer learning (Cheng & Malhi, 2017, Liu et al., 2017) and generative data augmentation approaches (Al-Dhabyani et al., 2019, Fujioka et al., 2019) have been proposed.

Classification approaches One commonly used ML approach in US CAD systems is the direct classification of the images collected by medical experts (Han et al., 2017, Meng et al., 2017). Indeed, different studies adopted deep learning classification approaches to identify various pathologies such as tumors in breast ultrasound (Tanaka et al., 2019, Becker et al., 2018, Y. Wang et al., 2020), liver pathologies (Acharya et al., 2015, Meng et al., 2017), thyroid nodules (Liu et al., 2017, Song et al., 2019), and others (Akkus et al., 2019).

Segmentation and detection approaches Detection and segmentation techniques designed to extract Regions of Interest (ROI) in US images are also common. For example, one solution extracts a ROI of the femoral cartilage from US images using segmentation (Kompella et al., 2019). Other approaches rely on object detection architectures to detect multiple ROIs within a US image for example to detect and classify breast lesions (Cao et al., 2019) or to detect different types of diseases in several organs (Zeng et al., 2020). Another example is SonoNet (Baumgartner et al., 2017), a real-time detection network that identifies fetal standard scan planes in ultrasound 2D images. A method combining segmentation and classification has been proposed for rheumatoid arthritis (Hemalatha et al., 2019). Specifically, the work aims at segmenting the synovial region in US images of metacarpophalangeal and proximal interphalangeal joints. The objective is to classify the grade of fluid expansion in the synovial region.

Multi-task learning approaches Previous works have explored the multi-task combination of classification and detection for non-US medical images (Yan et al., 2019, Gao et al., 2020, Sainz de Cea et al., 2020, Lin et al., 2017, Le et al., 2019). 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 which is divided into two different decoders for the two tasks (Gong et al., 2021). Zhang et al. adopt a multi-task learning algorithm to segment and classify cancer in Breast US images. They propose to use DenseNet121 as backbone, followed by a decoder branch with layers connected by attention-gated (AG) units to segment the images (Zhang et al., 2021). The second branch performs a classification task that takes in input the features extracted by the encoder. We are not aware of



(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*

Fig. 10. Detection examples. Green represents the ground truth, red and blue the results of the *Multi-Task approach* and *Detection approach*, respectively.

multi-task learning algorithms adopted to address the problem of joint recess distension detection, and more generally we found no prior works proposing multi-task networks to analyze musculoskeletal US images.

6.2. Classification and detection of joint recess distension

US images are commonly used for joint assessment and detection of joint recess distension in hemophilic patients. For this task, HEAD-US (Martinoli et al., 2013) is a standardized protocol to support physicians in acquiring US images of commonly affected joints and formulating a diagnosis.

Different solutions have been proposed in the literature to automatically detect and classify joint recess distension. For instance, a CNN-based method has been proposed to perform segmentation and classification of the Bicipital Peritendinous Effusions on the shoulder joint (Lin et al., 2020). Specifically, a VGG-16 (Simonyan & Zisserman, 2014) network is used for feature extraction, and a second CNN is used to classify the distension in three classes (*i.e.*, mild, moderate, and severe). The authors evaluated their method on a dataset of 3801 images, including both healthy individuals and individuals with BPE with various severity levels, reaching an accuracy of 75%.

Another work considers the knee joints (Long et al., 2020) and uses segmentation techniques to classify different types of pathologies inside 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 that, similarly to our work, extracts Sub-Quadricipital Recess (SQR) of the knee joint from US images (Wang et al., 2022). 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. Differently from that work, besides identifying the SQR area, we also classify it into distended and not distended. Indeed, ARB U-Net only considers distended SQR US images.

There are two main differences between our paper and the three works mentioned above. First, they all adopt a segmentation based approach that requires the exact target area to be annotated by an expert practitioner. Our approach, instead, only requires the practitioner to annotate the SQR bounding box, which is much simpler and fast. The second difference is that the three papers above do not focus on patients with hemophilia.

A recent abstract paper Tyrrell et al. (2021) considers US images of patients with hemophila and addresses the problem of classifying distended and not-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%. However, that prior work does not describe the methodology used for the classification, and it does not perform detection, differently from our proposed solutions.

A direct quantitative comparison between previous works and our paper is not possible for two reasons. First, the datasets used for the evaluation of previous works are not public and hence we cannot eval-

uate our techniques with the data used in previous works. The second reason is that running existing solutions on our dataset is not possible neither, because the first three papers mentioned above require the recess segmentation mask, which we do not have, while the last one does not report sufficient details to reproduce the proposed solution.

7. Conclusion

Early detection of hemarthrosis is fundamental to reduce the risk of under and over treatment for hemophilic patients. A Computer-Aided Diagnosis (CAD) tool that detects joint recess distension from ultrasound (US) images can support practitioners in diagnosing hemarthrosis without the need for expensive and time consuming exams (like MRI). We investigate the requirements of such a tool 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, and there can even be borderline cases in which the recess is only partially distended. Finally, the datasets in this problem domain are generally small and may contain noisy images.

This paper presents two solutions, each providing both recess detection and classification. Experiments, conducted on images of the SQR (*i.e.*, the knee joint recess), focusing on a specific US scan (the SQR longitudinal scan) show promising results. Indeed, both solutions achieve a balanced accuracy of approximately 0.75, a threshold value used in the literature to distinguish "useful" medical tests (Power et al., 2013). In particular the *Multi-task approach* achieves a BA value of 0.78. For what concerns the detection, the two solutions guarantee a correct detection (*i.e.*, IoU > 0.5) in more than 82% of the cases.

We believe that the performance of our solutions can be considerably improved in two possible ways. First, the CAD tool could process multiple images of the same joint, possible from different scans or from a video feed. The different results computed on the various images can then be combined to provide a more reliable outcome. The second improvement could be the adoption of an ensemble approach, in which a number of different models are trained and the CAD tool computes the result as a combination of the individual results provided by each model. Beyond improving the performance, these possible improvements have a potential important advantage: they can identify borderline cases (*e.g.*, when there is a disagreement in classification by two or more models, or by processing two scans of the same knee). In these (hopefully rare) cases, the CAD tool can inform the practitioner who can decide, for example, to use a different diagnosing tool (*e.g.*, MRI).

As a future work we intend to apply the proposed solutions on multiple scans for multiple joints. Actually, we are currently collecting images and videos on a total of 6 scans for the knee, the elbow, and the ankle. Our final goal is to create a bed-side solution for early hemarthrosis diagnosis. The idea is to enable an operator with little training (*e.g.*, the patient or a caregiver) to acquire US images with a portable US device. The system will support the operator during the acquisition by identifying the relevant reference points (*e.g.*, the patella), by guiding the operator to correctly position the US probe, and by evaluating the images in real time to inform the operator that a suitable image was collected. The US images will then be transmitted to the practitioner or even automatically processed as a part of a screening procedure.

CRediT authorship contribution statement

Claudio Bettini: Supervision, Conceptualization, Methodology, Writing – Review & Editing, Funding acquisition, Project administration. **Flora Peyvandi:** Supervision, Conceptualization, Methodology, Writing – Review & Editing, Resources, Funding acquisition, Project administration. **Roberta Gualtierotti:** Conceptualization, Data collection, Methodology, Investigation, Data curation, Writing – Review & Editing, Resources. **Gabriele Civitarese:** Methodology, Validation, Investigation, Writing – Original Draft, Writing – Review & Editing. **Dragan Ahmetovic:** Methodology, Validation, Investigation, Writing – Original Draft, Writing – Review & Editing. **Sergio Mascetti:** Conceptualization, Methodology, Validation, Investigation, Writing – Original Draft, Project administration, Writing – Review & Editing, Funding acquisition. **Marco Colussi:** Software, Methodology, Validation, Formal Analysis, Investigation, Writing – Original Draft, Visualization, Data curation, Writing – Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was partially supported by Italian Ministry of University and Research with project "MUSA-Multilayered Urban Sustainability Action" (project ID: ECS_00000037), NextGeneration EU, PNRR and by Italian taxes "5 x 1000 - 2020" devolved to Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico.

References

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

Akkus, Z., Cai, J., Boonrod, A., Zeinoddini, A., Weston, A. D., Philbrick, K. A., & Erickson, B. J. (2019). A survey of deep-learning applications in ultrasound: Artificial intelligence-powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology*, *16*(9), 1318–1328.

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2019). Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. *International Journal of Advanced Computer Science & Applications*, *10*(5).

Baumgartner, C. F., Kamnitsas, K., Matthew, J., Fletcher, T. P., Smith, S., Koch, L. M., Kainz, B., & Rueckert, D. (2017). Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, *36*(11), 2204–2215.

Becker, A. S., Mueller, M., Stoffel, E., Marcon, M., Ghafoor, S., & Boss, A. (2018). Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: A pilot study. *The British Journal of Radiology*, *91*, Article 20170576.

Bochinski, E., Senst, T., & Sikora, T. (2017). 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.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934.

Brattain, L. J., Telfer, B. A., Dhyani, M., Grajo, J. R., & Samir, A. E. (2018). Machine learning for medical ultrasound: Status, methods, and future opportunities. *Abdominal Radiology*, *43*(4), 786–799.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124). IEEE.

Cao, Z., Duan, L., Yang, G., Yue, T., & Chen, Q. (2019). An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Medical Imaging*, *19*(1), 1–9.

Chan, H.-P., Charles, E., Metz, P., Lam, K., Wu, Y., & Macmahon, H. (1990). Improvement in radiologists' detection of clustered microcalcifications on mammograms. *Arbor*, *1001*. 48109–0326.

Chan, V., & Perlas, A. (2011). Basics of ultrasound imaging. In *Atlas of ultrasound-guided procedures in interventional pain management* (pp. 13–19). Springer.

Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2021). Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, *7*(4), 750–758.

Cheng, P. M., & Malhi, H. S. (2017). Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of Digital Imaging*, *30*(2), 234–243.

Ci, B., & Rule, R.-O. (1987). Confidence intervals. *Lancet*, *1*(8531), 494–497.

Doi, K. (2005). Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*, *78*(suppl_1), s3–s19.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fujioka, T., Mori, M., Kubota, K., Kikuchi, Y., Katsuta, L., Adachi, M., Oda, G., Nakagawa, T., Kitazume, Y., & Tateishi, U. (2019). Breast ultrasound image synthesis using deep convolutional generative adversarial networks. *Diagnostics (Basel)*, *9*(4), 176.

Fujita, H. (2020). Ai-based computer-aided diagnosis (ai-cad): The latest review to read first. *Radiological Physics and Technology*, *13*(1), 6–19.

Gao, F., Yoon, H., Wu, T., & Chu, X. (2020). A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications*, *143*, Article 112957.

Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., & Li, G. (2021). 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.

Gualtierotti, R., Solimeno, L. P., & Peyvandi, F. (2021). Hemophilic arthropathy: Current knowledge and future perspectives. *Journal of Thrombosis and Haemostasis*, *19*(9), 2112–2121.

Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., & Seong, Y.-K. (2017). A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine and Biology*, *62*(19), 7714.

Hemalatha, R., Vijaybaskar, V., & Thamizhvani, T. (2019). Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine*, *233*(6), 657–667.

Hilgartner, M. W. (2002). Current treatment of hemophilic arthropathy. *Current Opinion in Pediatrics*, *14*(1), 46–49.

Huang, Q., Zhang, F., & Li, X. (2018). Machine learning in ultrasound computer-aided diagnostic systems: A survey. *BioMed Research International*, *2018*.

Jabeen, K., Khan, M. A., Alhaisoni, M., Tariq, U., Zhang, Y.-D., Hamza, A., Mickus, A., & Damaševičius, R. (2022). Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*, *22*(3), 807.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K., Lorna, A. V., Montes, D., Nadar, L., Laughing, tkianai, y., Skalski, P., Wang, Z., Hogan, A., Fati, C., ... Minh, M. T. (2022). ultralytics/yolov5: v6.1 - tensorrt, tensorflow edge tpu and openvino export and inference. https://doi.org/10.5281/zenodo.6222936.

Joshua, F., Lassere, M., Scheel, A. K., Kane, D., Grassi, W., Conaghan, P. G., Wakefield, R. J., D'Agostino, M.-A., Bruyn, G. A., Szkudlarek, M., Naredo, E., Schmidt, W. A., Balint, P., Filippucci, E., Backhaus, M., & Iagnocco, A. (2007). Summary findings of a systematic review of the ultrasound assessment of synovitis. *Journal of Rheumatology*, *34*(4), 839–847.

Kompella, G., Antico, M., Sasazawa, F., Jeevakala, S., Ram, K., Fontanarosa, D., Pandey, A. K., & Sivaprakasam, M. (2019). Segmentation of femoral cartilage from knee ultrasound images using mask r-cnn. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 966–969).

Le, T.-L-T., Thome, N., Bernard, S., Bismuth, V., & Patoureaux, F. (2019). Multitask classification and segmentation for cancer diagnosis in mammography. ArXiv preprint, arXiv:1909.05397.

Lin, B.-S., Chen, J.-L., Tu, Y.-H., Shih, Y.-X., Lin, Y.-C., Chi, W.-L., & Wu, Y.-C. (2020). Using deep learning in ultrasound imaging of bicipital peritendinous effusion to grade inflammation severity. *IEEE Journal of Biomedical and Health Informatics*, *24*(4), 1037–1045.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, T., Xie, S., Yu, J., Niu, L., & Sun, W. (2017). 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.

Long, Z., Zhang, X., Li, C., Niu, J., Wu, X., & Li, Z. (2020). Segmentation and classification of knee joint ultrasonic image via deep learning. *Applied Soft Computing*, *97*, Article 106765.

Martinoli, C., Alberighi, O. D. C., Di Minno, G., Graziano, E., Molinari, A. C., Pasta, G., Russo, G., Santagostino, E., Tagliaferri, A., Tagliafico, A., & Morfini, M. (2013). Development and definition of a simplified scanning procedure and scoring method for haemophilia early arthropathy detection with ultrasound (head-us). *Thrombosis and Haemostasis*, *109*(6), 1170–1179.

Meng, D., Zhang, L., Cao, G., Cao, W., Zhang, G., & Hu, B. (2017). Liver fibrosis classification based on transfer learning and fcnet for ultrasound images. *IEEE Access*, *5*, 5804–5810.

Mohiyuddin, A., Basharat, A., Ghani, U., Abbas, S., Naeem, O. B., & Rizwan, M. (2022). Breast tumor detection and classification in mammogram images using modified yolov5 network. *Computational & Mathematical Methods in Medicine*, *2022*.

Peyvandi, F., Garagiola, I., & Young, G. (2016). The past and future of haemophilia: Diagnosis, treatments, and its complications. *The Lancet*, *388*(10040), 187–197.

Plut, D., Kotnik, B. F., Zupan, I. P., Kljucevsek, D., Vidmar, G., Snoj, Z., Martinoli, C., & Salapura, V. (2019). Diagnostic accuracy of haemophilia early arthropathy detection with ultrasound (head-us): A comparative magnetic resonance imaging (mri) study. *Radiology and Oncology*, *53*(2), 178–186.

Power, M., Fell, G., & Wright, M. (2013). Principles for high-quality, high-value testing. *BMJ Evidence-Based Medicine*, *18*(1), 5–10.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. ArXiv.org.

Roosendaal, G., & Lafeber, F. P. J. G. (2003). Blood-induced joint damage in hemophilia: Modern management of hemophilia a to prevent bleeding and arthropathy. *Seminars in Thrombosis and Hemostasis*, *29*(1), 37–42.

Sainz de Cea, M. V., Diedrich, K., Bakalo, R., Ness, L., & Richmond, D. (2020). 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.

Sarvamangala, D., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: A survey. *Evolutionary Intelligence*, *15*(1), 1–22.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician*, *55*(3), 182–186.

Sharma, N., Jain, V., & Mishra, A. (2018). An analysis of convolutional neural networks for image classification. *Procedia Computer Science*, *132*, 377–384.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv:1409.1556.

Song, J., Chai, Y. J., Masuoka, H., Park, S.-W., Kim, S-j., Choi, J. Y., Kong, H.-J., Lee, K. E., Lee, J., Kwak, N., et al. (2019). Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine*, *98*(15).

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

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning, PMLR* (pp. 1139–1147).

Tanaka, H., Chiu, S.-W., Watanabe, T., Kaoku, S., & Yamaguchi, T. (2019). Computer-aided diagnosis system for breast ultrasound images using deep learning. *Physics in Medicine and Biology*, *64*(23), Article 235013.

Tingelhoff, K., Eichhorn, K. W., Wagner, I., Kunkel, M. E., Moral, A. I., Rilk, M. E., Wahl, F. M., & Bootz, F. (2008). Analysis of manual segmentation in paranasal ct images. *European Archives of Oto-Rhino-Laryngology*, *265*(9), 1061–1070.

Tyrrell, P., Blanchette, V., Mendez, M., Paniukov, D., Brand, B., Zak, M., & Roth, J. (2021). 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. *Research and Practice in Thrombosis and Haemostasis*, *5*.

Tzutalin, L. (2015). https://github.com/tzutalin/labelImg.

Wang, C.-Y., Mark Liao, H.-Y., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). 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).

Wang, Y., Choi, E. J., Choi, Y., Zhang, H., Jin, G. Y., & Ko, S.-B. (2020). Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. *Ultrasound in Medicine & Biology*, *46*(5), 1119–1132.

Wang, Z., Yang, Q., Liu, H., Mao, L., Zhu, H., & Gao, X. (2022). 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.

Wells, P. N. T. (2006). Ultrasound imaging. *Physics in Medicine and Biology*, *51*(13), R83–R98.

Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., & Summers, R. M. (2019). 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.

Zeng, X., Wen, L., Liu, B., & Qi, X. (2020). Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*, *392*, 132–141.

Zhang, G., Zhao, K., Hong, Y., Qiu, X., Zhang, K., & Wei, B. (2021). 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*, *16*(10), 1719–1725.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). 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).