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

Li, Y.L.; Hassanzadeh, T.; Shamonin, D.P.; Reijnierse, M.; Mil, A.H.M.V.; Stoel, B.C.

Biomedical Signal Processing and Control 91 (2024) 105990

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

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

Yanli Li [a], Tahereh Hassanzadeh [a], Denis P. Shamonin [a], Monique Reijnierse [b], Annette H.M. van der Helm-van Mil [c], Berend C. Stoel [a,*]

[a] Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[b] Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[c] Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands

ARTICLE INFO

Keywords:
Rheumatoid arthritis
Deep learning
Self-supervised learning
Wrist MRI

ABSTRACT

Predicting the development of rheumatoid arthritis (RA) in an early stage through magnetic resonance imaging (MRI) can initiate timely treatment and improve long-term patient outcomes. Although manual prediction is time-consuming and requires expert knowledge, automatic RA prediction has not been fully investigated. While standard models fail to achieve acceptable performance, we present a consistency-based deep learning framework to classify and predict RA automatically and precisely, including an output-standardized model, customized self-supervised pretraining and a loss function that is based on label consistency between original and augmented inputs. For training and evaluation, we used a database, containing 5945 MRI scans of carpal, metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints, from 2151 subjects obtained over a period of ten years. Four (three classification- and one prediction-) tasks were defined to distinguish two patient groups (with recent-onset arthritis and clinically suspect arthralgia) from healthy controls and RA from other arthritis patients within the recent-onset arthritis group, and predict RA development in a period of two years within the clinically suspect arthralgia group. The proposed method was evaluated with the area under the receiver operating curve (AUROC) on a separate test set, achieving mean AUROCs of 83.6%, 83.3%, and 69.7% in the three classification tasks, and 67.8% in the prediction task. This proves the existence of early signs of RA in MRI and the potential of a consistency-based deep learning model to detect these early signs and predict RA.

1. Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disorder that especially affects joints in wrists, hands, and feet [1]. It can ultimately result in bone erosions and joint deformations and only very early detection and treatment can improve the long-term outcome [2]. Finding early signs, localizing lesions, and predicting potential development into RA can help radiologists and rheumatologists to diagnose and treat RA at an early stage. Therefore, this motivated our study to find early RA-relevant signs through imaging. Magnetic resonance imaging (MRI), which enables the visualization of both anatomical information and inflammatory signs, is the most sensitive imaging method to detect inflamed areas and has become a common imaging modality for RA research. RAMRIS (rheumatoid arthritis magnetic resonance imaging scoring system) [3] is currently the most widely-used imaging biomarker to quantitatively score RA for each anatomical site [4–6]. To classify and predict early inflammatory signs, RAMRIS assesses bone marrow edema [5], synovitis [7], and tenosynovitis. However, scoring these biomarkers is time-consuming, requires expert training, and depends on prior knowledge and assumptions to detect early signs.

In previous work, automated biomarker quantification methods were proposed and demonstrated a high correlation with expert RAMRIS scores [4]. These pre-defined image features may, however, not be the optimal biomarkers to classify and predict RA. Moreover, certain inflammatory signs may not be relevant to RA, as they also appear in healthy individuals. Therefore, the visual scoring by RAMRIS also compares with healthy controls. This makes it challenging to classify and predict RA through traditional image analysis methods.

Since these tasks typically include labeling or classification, deep learning (DL) methods are highly suitable, without relying too much on prior assumptions or pre-defined imaging biomarkers. Despite the success in other medical imaging labeling tasks [8–10], DL methods

* Corresponding author.
E-mail addresses: y.li2@lumc.nl (Y. Li), t.hassanzadeh@gmail.com (T. Hassanzadeh), d.p.shamonin@lumc.nl (D.P. Shamonin), M.Reijnierse@lumc.nl (M. Reijnierse), a.h.m.van_der_helm@lumc.nl (A.H.M. van der Helm-van Mil), b.c.stoel@lumc.nl (B.C. Stoel).

https://doi.org/10.1016/j.bspc.2024.105990
Received 8 November 2023; Received in revised form 8 January 2024; Accepted 29 January 2024
Available online 3 February 2024

have not been fully investigated yet in the classification and prediction of RA due to the following reasons. Firstly, the time window is narrow for collecting images from arthralgia patients with possible early signs of RA, complicating data acquisition and resulting in a limited dataset size. Consequently, overfitting becomes a severe problem, and the size of the dataset also restricts the performances of large models with massive trainable parameters. Secondly, the variety and complexity of anatomical and pathological structures in hands and feet, and variability in positioning hands and feet, further amplifies the difficulty of the tasks, resulting in insufficient performances of standard models for medical images. Thirdly, artifacts caused by fat suppression errors, movement or aliasing may significantly influence the automatic interpretation of MRI scans. These artifacts may appear more often in certain time periods, becoming serious confounders while splitting the dataset for evaluation, and worsening the overfitting problem. Finally, there are no publicly accessible datasets for similar subjects or tasks, therefore DL models cannot benefit from transfer learning and well-developed pre-processing methods.

To overcome these challenges and predict RA in an early stage, we propose a so-called consistency-based training framework for a simple deep learning model to pre- and post-process MRI scans, and predict the development of RA in patients with recent-onset arthritis or clinically suspect arthralgia. This consistency-based framework helps the model to utilize the unchanged information and learn from a limited number of samples. Specifically, we first pre-trained the model with a self-supervised reconstruction method, based on a masked autoencoder (MAE) [11], to let the model understand the anatomy of human hands and feet by filling in the masked areas in MRIs from the training set. Meanwhile, we applied an extra contrastive loss function based on augmentation to emphasize that the disease-related information should be invariant to spatial transformations (i.e., the output probabilities or logits should be independent of an object's position or orientation).

Main contributions of this paper are: (1) This is the first MRI-based early RA prediction framework using deep learning with promising results; (2) A self-supervised reconstruction is applied for pre-training to utilize the anatomical consistency of human hands and feet, thereby replacing Transformers [12] by fully convolutional networks (FCNs) that have far less parameters than the visual learner [11]; (3) A contrastive loss function is defined to accelerate the training process and force the model to focus on unchanged RA information after augmentation.

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

2. Dataset and task design

2.1. Structure of materials

The models were trained and evaluated based on a database (informed consent given by all patients, LUMC protocol reference number: B19.008 and P11.210) that contained a total of 5945 MRI scans of carpal, metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints, from 2151 subjects obtained over a period of ten years (see Fig. 1). This MRI dataset consists of three groups: 1247 patients with recent-onset arthritis, called early arthritis clinic (EAC), 727 arthralgia patients with an increased risk of developing RA, called clinical suspect arthralgia (CSA), and 177 healthy controls as atlas (ATL). Study protocols for the EAC cohort (reference number: B19.008, date: 29-may-2009) and CSA/healthy controls(ATL) cohort (reference number: P11.210, date 08-feb-2012) were approved by the local Medical Ethical Committee of the Leiden University Medical Center (LUMC).

The EAC group consists of patients with clinically confirmed arthritis, of which a subgroup was diagnosed with RA within a year, whereas the remainder was diagnosed with other arthritides (non-RA) or undifferentiated arthritis (UA). According to these diagnoses after one year, EAC patients were divided into either RA or non-RA/UA, indicated by EAC(RA+) and EAC(RA-). The classification task was to distinguish these two subgroups. The CSA group was followed over a period of two years in order to establish whether they had developed RA. The CSA group was divided into two groups CSA(RA+) and CSA(RA-), with the task to distinguish these two subgroups, so as to predict the development of RA. The ATL group was collected over a shorter time period. Further details (including patients' characteristics) of the collected dataset can be found in [4].

In each group, the carpal, MCP, and MTP joints were scanned with a 1.5T extremity MRI scanner (GE Healthcare) using a 100-mm coil, with contrast enhancement (T1-Gd) and frequency-selective fat saturation. For coronal scans (3D scans with the highest resolution in the coronal plane), the repetition time was 650 ms, echo time 17 ms, acquisition matrix 364 × 224, echo train length 2, slice thickness 2 mm, and slice gap 0.2 mm. For transversal (axial) scans, these parameters are: 570 ms, 7 ms, 320 × 192, 2, 3 mm, and 0.3 mm, respectively [4]. The scans were reconstructed into [512, 512, 20±5] images, which means the resolution in the Z direction was relatively low, leading to information loss and thus increasing the importance of fusing information from coronal and axial scans.

2.2. Task definition

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

3. Methods

3.1. Overall workflow

Fig. 2 presents the overall workflow and basic information of the proposed methods for training the CNNs for the four tasks. The MRI scans from different anatomical sites were processed by a unified process. The process begins with preprocessing to standardize the output, removing background noise and artifacts, resizing the anatomical structures in the images into a fixed size, slice-by-slice intensity normalization and selecting the central slices for axial scans to increase the information density. This is followed by a simple model pre-trained through self-supervised reconstruction as the feature extractor to obtain RA-related features for DL interpretation. Trained by comparing with the true label and the so-called 'label consistency' between the original and augmented image, the classification or prediction result for a specific task is produced as output.

The next four subsections will successively introduce the preprocessing, the backbone model, the self-supervised pretraining, based on the anatomical consistency, and the contrastive loss function, based on label consistency of the same samples.

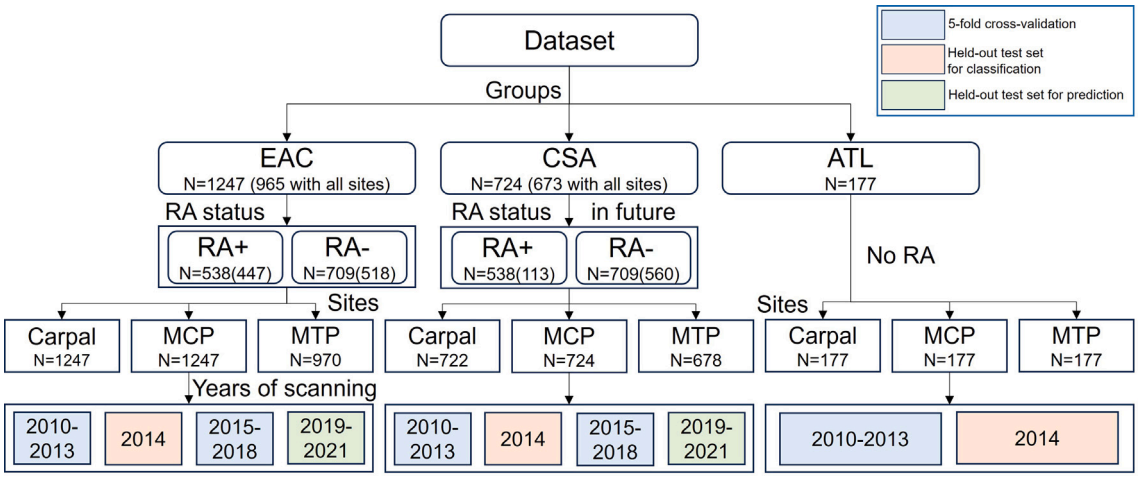

Fig. 1. Dataset composition, task definition, and evaluation design. The dataset consists of three major groups: patients with recent-onset arthritis (EAC), clinically suspect arthralgia (CSA) and healthy controls (ATL). The EAC group was divided into EAC(RA+) and EAC(RA-), while the CSA group was divided into CSA(RA+) and CSA(RA-), where "RA+"/ "RA-" indicates the RA status one or two years after the baseline. Each group contains MRI scans collected from the carpal (wrist), metacarpophalangeal (MCP), and metatarsophalangeal (MTP) joints. The dataset was collected between 2010 and 2021, while the ATL group was collected over a shorter time period (between 2010 and 2014).

Table 1
Description of the four tasks.

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

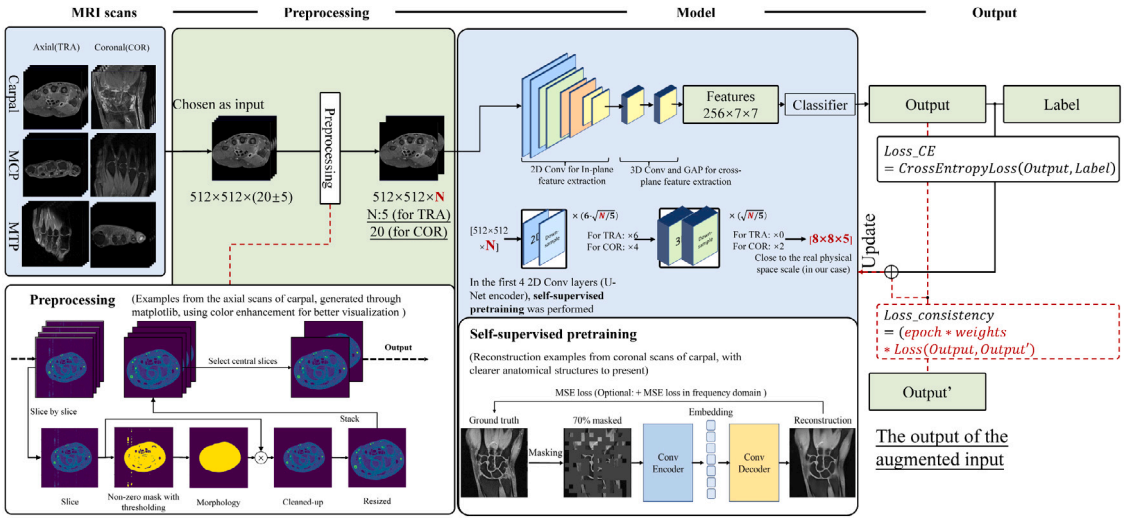

Fig. 2. Overall workflow. The training framework includes three main steps: (a) Preprocessing; (b) Model and its self-supervised pretraining; and (c) a self-contrastive loss function. COR: coronal; TRA: transversal (axial); MSE: mean squared error; Conv: convolutional layer. The value N, which represents the number of slices, is set to be five for TRA scans and twenty for COR scans to improve the information density based on our previous study [13,14].

3.2. Preprocessing

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

Images were first thresholded at 10% of the maximum intensity value in the image and processed through morphological opening and closing operations [15] to obtain the masks of targeted foreground objects [16]. The threshold was fixed according to a small subgroup of this dataset, and visually checked on the training set. Some thresholding algorithms such as OTSU might improve the thresholding process and help to generalize the preprocessing to other datasets as some studies stated the OTSU outperforms fixed thresholding [17]. However, in this study, fixed thresholding outperformed OTSU in distinguishing the foreground and background in most cases. The automatic thresholding methods could cause information loss due to over-thresholding in the targeted anatomical structures (some examples were presented in the

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

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

Moreover, the size and foreground-to-background ratio of 3D MRI scans reduced the efficiency and amplified the difficulty of model training. Therefore, based on previous work [13], and the observation that the foreground-to-background ratio decreases significantly with increasing distance from the central slice, the central five slices were selected as input instead of the whole 3D scans. Here, the central five slices were defined as the slices with the largest sum of non-zero mask area in the previous masking process. The number of central slices (N in Fig. 2) was determined based on pre-experiments from our previous study [14], in which five central slices could perform as well as using all 3D scans in TRA. For the coronal (COR) scans, the variety of spatial placements and the irregular gradient intensity change in the scans make it difficult to select the central slices automatically. Therefore, in this study, the N for TRA scans is set to five and for COR is set to twenty.

3.3. Backbone model

Because the COR and TRA scans describe the same anatomical sites (carpal, MCP, and/or MTP joints), but with different resolutions in each direction, the model architecture was designed to adapt different sizes of images from TRA and COR scans and then output the extracted features in a fixed shape. Considering the limited number of samples and different task complexities, we implemented a model transferred from the basic U-Net [18] encoder as the backbone, for potential pretraining and transfer learning. The model architecture contains two main parts: (1) an encoder that contains both 2D and 3D convolutional (Conv) layers to output features of a same size for different scans; and (2) a standard dense layer as a classifier.

The encoder is formed by sequentially stacking 2D (kernel size:[1, 3, 3]) and 3D (kernel size:[3, 3, 3]) Conv layers, with the same hyperparameters as the basic U-Net. Inspired by the resampling process in nnU-Net [16] that standardizes the input size at the image level, the number of 2D and 3D Conv layers is set to accommodate samples with different input sizes and output a fixed number of features to perform a feature-level standardization. The reasons for using the U-Net encoder, with a few changes, as the backbone of 2D Conv parts are: (1) Most advanced model architectures and functional modules require a large amount of training data, which is not available; (2) U-Net encoders are simple to be implemented and reproduced for both researchers and users; (3) U-Net encoders, which have been widely used in both natural (known as the VGG encoder) and medical imaging field (encoder part of a U-Net), is naturally convenient for pretraining through transfer learning or self-supervised training (see next section).

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

3.4. Anatomical consistency and self-supervised pretraining

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

To pre-train the model when samples are limited, a self-supervised method called masked auto-encoder (MAE) [11] was employed. Compared to natural images analyzed by self-supervised methods, our number of samples is limited, yet the structures of human hands and feet are anatomically consistent, which could be learned by models and used as prior knowledge. For example, the number of bones or the existence of inflammation around tendons in the wrist could be a hint to finding bone marrow edema and tenosynovitis, respectively, that are related to RA. In a previous study, a self-supervised pretraining strategy was explored in medical imaging by [19] with a series of augmentation methods, proving the potential of applying self-supervised reconstruction as pretraining to 'warm up' the model with a similar task. Compared to the method of Zhou, which requires models to reconstruct original images through differently-augmented images, the training strategy of MAE is simpler, and more efficient by reconstructing original images from randomly-masked images. Since valid results with good generalization ability have been achieved by MAE in natural imaging fields, we extended it from a process based on Transformers [12] to a process based on the U-Net that contains less parameters, which is more suitable for medical imaging, but with the same principle of reconstruction from masked images to learn underlying semantics.

Fig. 3 presents the basic idea of building a self-supervised pretraining process on the U-Net encoder. To define the reconstruction task, 70% of the input image was masked by patches (16 × 16 pixels), which were distributed randomly. Using skip connections in this reconstruction-based pre-training would introduce a risk of having the model not learn the underlying patterns or anatomical structures at a high level, but copy and paste over the epochs to get the reconstruction in the corresponding areas. Therefore, we expected that the encoder would learn more underlying patterns of anatomical structures of human hands and feet, if skip connections are removed, to avoid information leakage through these shortcuts. Although models with skip connections could performed quite well on the training set, pre-experiments [14] showed that they failed to achieve good reconstructions in validation sets that are not involved in the training process.

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

3.5. Label-consistency loss function

Similar to other DL methods in medical imaging, basic data augmentation was applied to overcome overfitting. Besides classical ways of augmentation, we took a different approach to use data augmentation,

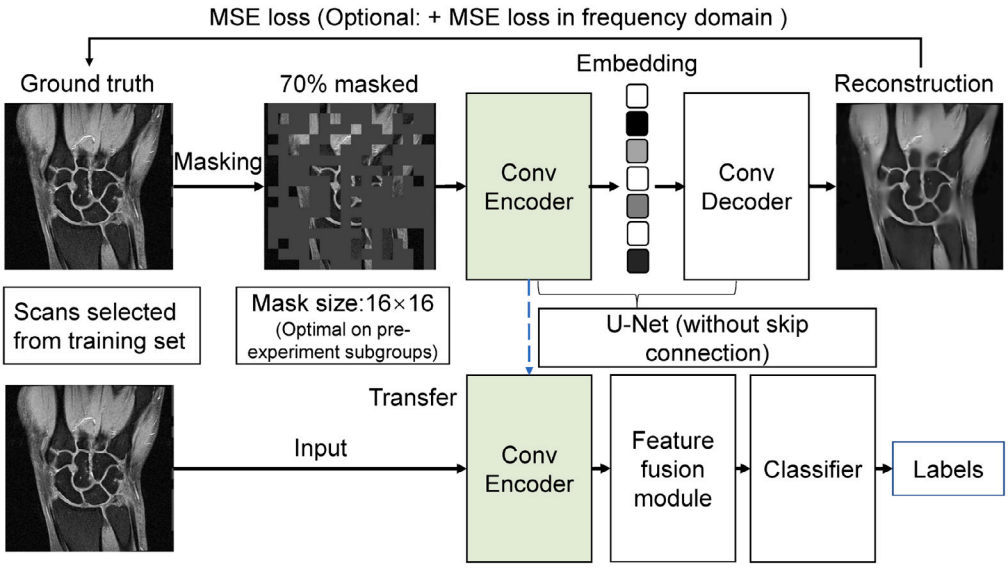

Fig. 3. The workflow of the self-supervised reconstruction. The mean squared error (MSE) loss contains two parts, the MSE loss of spatial domain and frequency domain.

inspired by contrastive learning [20–22], to maximize the value of the shared information in the original image and the augmented one. This could help model training to focus on the unchanged RA-related information, by excluding the impact of spatial placement and added noise more efficiently.

We propose a loss function, called the 'label-consistency' loss function, to take advantage of the fact that the properly-augmented and original input data share the same label and disease severity, which is the so-called "label consistency" of RA information. The assumption for this loss function is that the RA-related information remains unchanged during the augmentation, as a defined spatial transformation and added Gaussian noise will not remove any lesions or anatomical structures. To comply with this assumption, the augmentation is limited to avoid large transitions or extensive cropping, and a margin (0.05) for the loss between the output is set to leave some space for accidental cut-offs by augmentation. The consistency loss function is aiming at minimizing the differences in the output logits between the input and augmented image, in one training epoch, while the cross-entropy loss function is trying to maximize the differences between different classes. To stabilize the training process, an epoch-dependent weight is added to minimize the impact of this extra loss function at the early stage of training and increase the impact at a later stage. The consistency loss function is given by: $loss = CrossEntropy(output, GT) + i \times w \times Margin(MSE(output, output_a))$. The cross-entropy part of the equation is the same as in standard classification tasks, where $output$ represents the output logit of the model, and GT refers to the ground truths. The second part represents the consistency loss, where i refers to the index of the training epoch in the range from zero to the maximum number of epochs-1; and w is a hyper-parameter used to control the weight of consistency loss. At the last epoch, the product of i and w will reach 1.0 to gain an equal effect as the cross-entropy loss. This gradually rising weight is to avoid affecting the direction of model learning in the early stages of training and enabling the model to converge. In our experiments, the range of w was set from 0.005 to 0.01 as the total number of epochs varies from 100 to 200 because of task differences; $output$ and $output_a$ refer to the output logits of the model with as input the original image and its augmented version, respectively. The $Margin()$ function leaves space for small values of MSE loss that might be caused by accidental occlusion of information by augmentation. Fig. 4 presents the workflow after adding the loss function.

3.6. Class activation mapping

The class activation mapping (CAM) technique [23–25] is one of the most common techniques to open the deep learning black box. Since in our case, classification mainly applies to the center of the images, we applied the pixel-to-pixel calculation of the original gradients instead of the average gradients in Grad CAM. Moreover, to fully reflect the model's judgment criteria, we retained the negative parts, which are usually removed in standard CAM methods, in which only the positive regions are presumed to represent objects appearing in the background. As the regions with negative values in saliency maps can represent normal objects that decrease the confidence of reporting early RA, we preserved them, resulting in activation values in the background (air) greater than zero, but still representing "no contribution". Consequently, regions with activation values lower than that of air represent a negative contribution to the targeted label. The results of CAMs can be found in the next section, which illustrates the focus of the models for RA classification/prediction.

4. Results

4.1. Evaluation principles

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

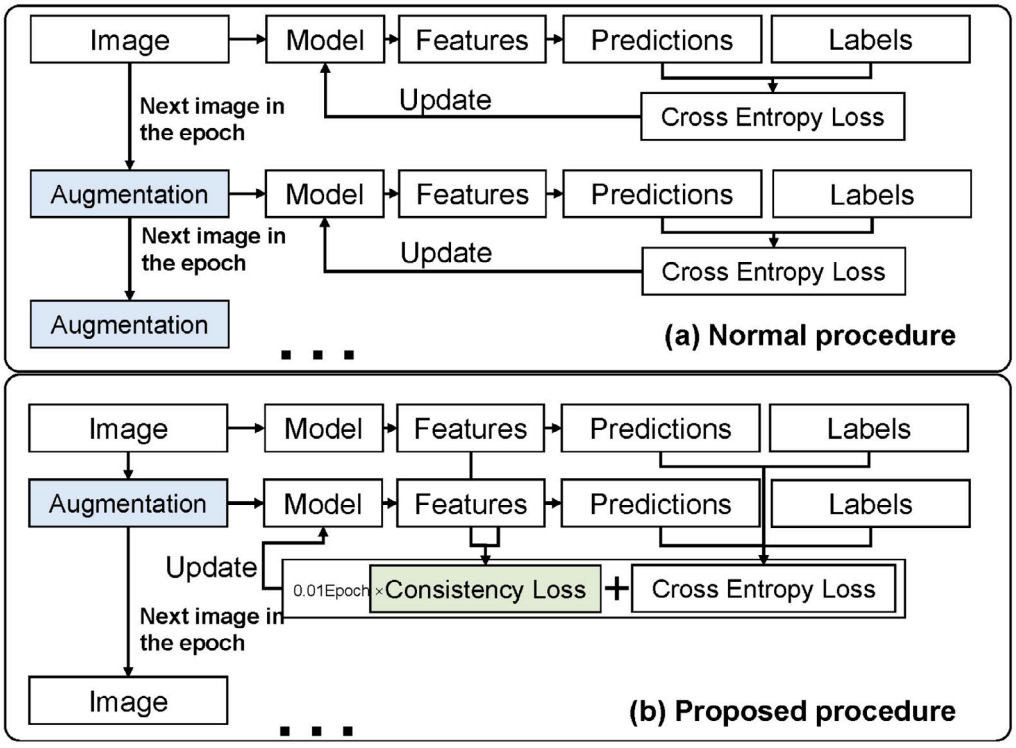

Fig. 4. (a) The normal procedure of using augmentation and (b) the procedure when the consistency loss function is added to the training process.

4.2. Reconstruction examples from self-supervised pretraining

Fig. 5 provides some examples from the self-supervised pre-training process. The first four convolutional layers, with a similar structure as the U-Net encoder, were trained to extract features from 70% masked images and reconstruct the original images with a decoder. As shown in Fig. 5(a), the model can predict unseen anatomical structures in the masked regions (highlighted by the red boxes) when 70% of the image in the test set was masked, without any labels or prior knowledge. When the 99%-masked images were fed into the models trained on 70%-masked images, as shown in Fig. 5(b), the model still grasped the basic concepts and structures of carpals although it was unable to predict the whole anatomy due to insufficiency of information. These results prove that the pre-trained models have learned some basic anatomical knowledge of carpals without any label.

4.3. Overall performance on the four tasks

The overall performance of the proposed method on all four tasks can be found in Fig. 6. The details of the input and results can be found in Table 2. The first two classification tasks served as pretraining for the RA classification/prediction tasks. In this phase, the proposed models were trained on the carpals, MCP and MTP, separately, and a combination of these scans. Due to the overfitting problem of MTP-based models, the combination of three anatomical sites failed to reach competitive results, therefore these are not presented in Fig. 6 and Table 2. Consequently, the models for MTP-based RA prediction were pre-trained by the reconstruction models only.

For classification tasks that distinguish early-onset arthritis or clinically suspect arthralgia from healthy controls, the models achieved AUROCs of over 0.8 on cross-validation and close level on the held-out test set. However, the performance dropped from around 0.65 to 0.7 for the third classification task. This mainly originates from the difficulty of distinguishing RA from other arthritides, as inflammatory areas may significantly contribute to distinguishing arthritides, yet are common in both RA and other arthritides. For the prediction task, which is more challenging, the performance of the DL model is comparable to statistical analysis on clinical variables [26] (AUROC equal to 0.74, with different validation set.).

Table 2
Overall performance on each task with different inputs.

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

4.4. General improvements compared to baseline models

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

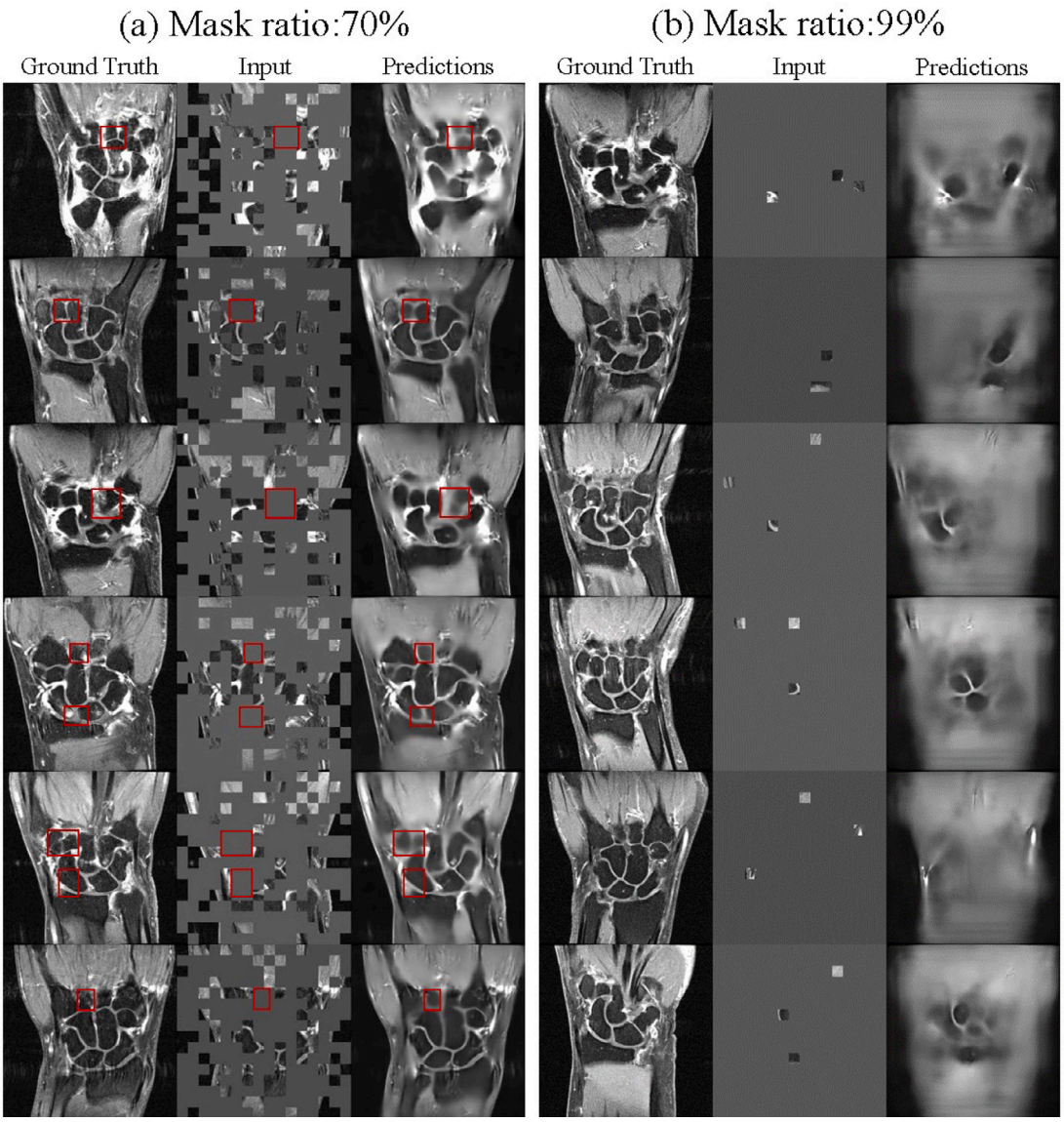

Fig. 5. Reconstruction examples from the model trained on 70% masked input, (a) test on 70% masked scans from the held-out test set. (b) test on 99% masked scans from the held-out test set.

with or without the attention module and dense block as the baseline for these four RA classification/prediction tasks. These models have been the most widely used backbone architectures for DL-based medical tasks. For example, in breast MRI [27], Alzheimer's MRI classification [28] and disc degenerative disease based on MRI [29], when LSTM and some other structures or modules (e.g. attention modules [30]) were introduced because of specific data characteristics, the CNN backbones remained to be ResNets and VGGs. Meanwhile, the comparison with these baseline models could more clearly prove the effectiveness of the proposed strategies.

ResNet3D in [31], which is the closest study to our task, was also implemented. However, the input resolution and tasks were different, making it fail to perform these tasks and cannot outperform the backbone (ResNet). For more advanced models, like Transformer, we were not able to train models because they are too data-hungry. Apart from the baseline models, we also applied the widely-used lightweight models such as MobileNet [32] and MobileViT [33] to investigate different types of models, the results can be found in the supplementary materials. Similar pre-experiments were also implemented to validate the effectiveness of other modules such as attention modules, multi-scale processing and multi-task training. However, all these attempts presented no statistical significance in improving the model performance.

As shown in Fig. 7, compared to all the baseline CNN models and a ViT-B [36] model, our models present substantial increases of AUROCs in the RA classification/prediction tasks. Especially in the CSA-related tasks, our models achieve significant AUROC improvements over 10% are achieved. Meanwhile, the MRI scans of carpals and MCPs appeared to be the most informative for the RA-related tasks. More details of the best models can be found in Table 3 and the performance of lightweight models can be found in the supplementary materials.

4.5. Ablation study of each proposed component

Fig. 8 presents the results of ablation experiments applied on the classification Task 1 and 2, which contain all the proposed strategies in the training process, based on the MRI of carpals. The self-supervised pretraining and consistency loss function contributed most to the performance, especially for the CSA classification task, while the model architecture and pre-processing also have a clear impact on the AUROC. The contribution of each strategy varies because of the variation in input materials, yet delivered a clear message that the performance of deep learning models in medical fields is highly dependent on the training strategies.

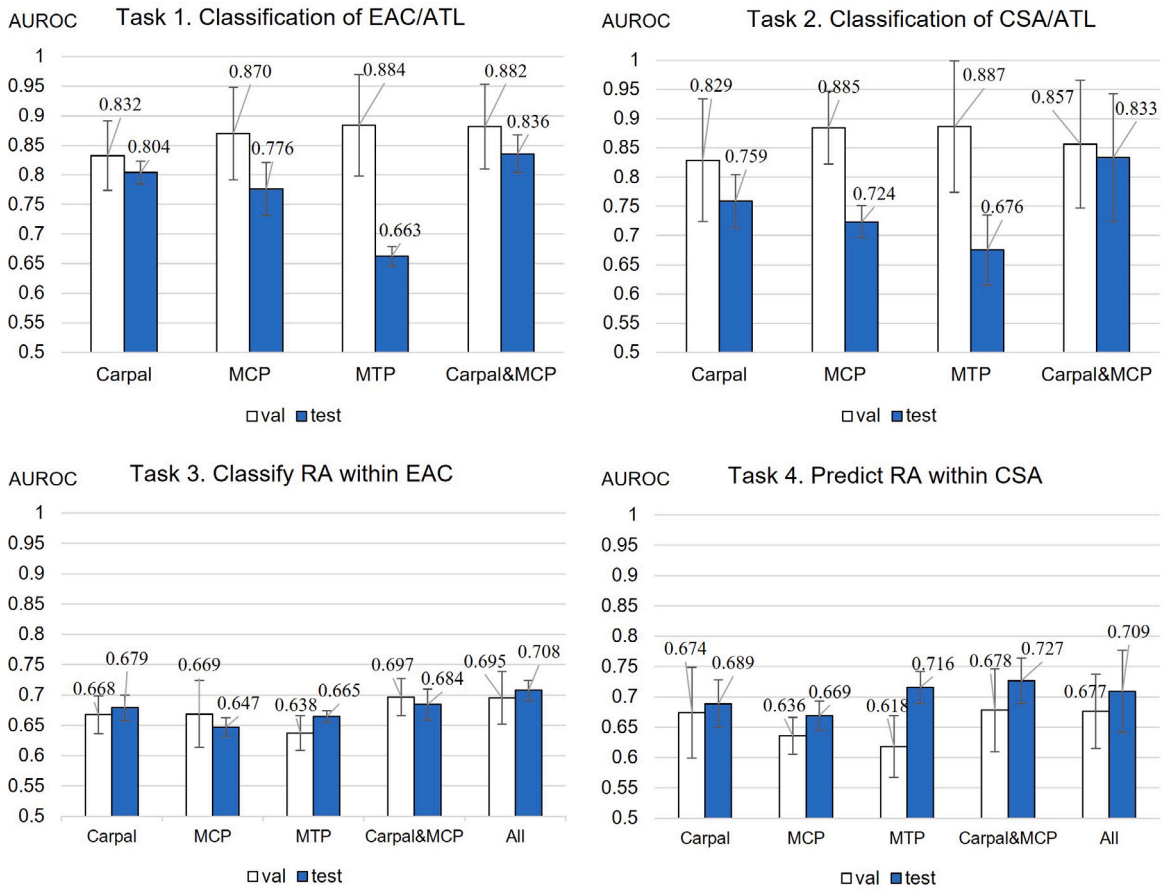

Fig. 6. Overall performance on the four tasks. The solid bars represent the results of the test set, while the hollow bars represent the results of the cross-validation. In most tasks, the ensemble models using the combination of carpals and MCPs obtained the highest AUROCs. The MTP-based models suffered overfitting and performed poorly on the EAC and CSA classification tasks, because of confounders related to the stability of the MRI scanner over time, which were analyzed in the supplementary materials.

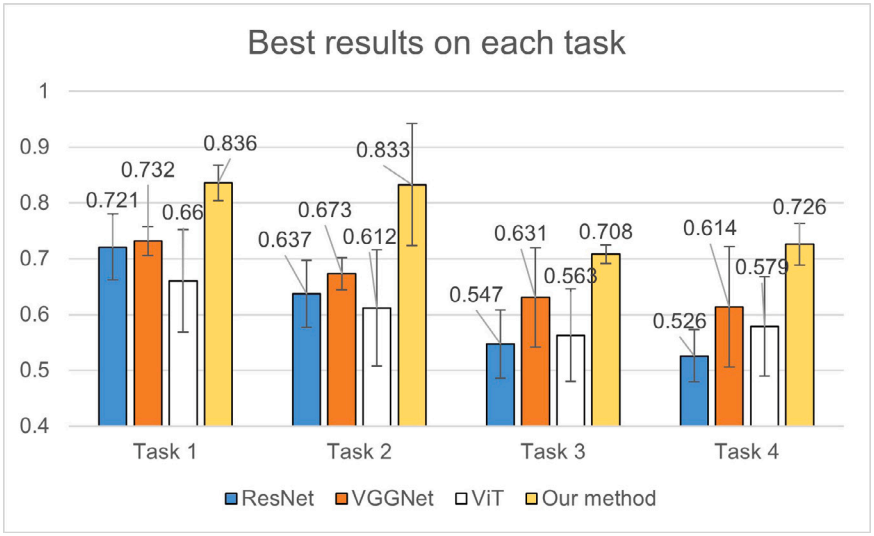

Fig. 7. Best results from each method applied on the test set. The blue bars present the results of ResNet34 [34] on test set in each task, while the orange bars are the results of VGG16 [35]. The best result of Transformer-like models is presented in white bars, using a ViT [36]. These models achieved the best performance of their kinds (ResNets/VGGs) in the four tasks. The results of the VGG models trained with the consistency-based methods (consistVGG16) are given by yellow bars.

4.6. Saliency maps generated by CAM

In Fig. 9, examples were randomly selected and organized based on the output confidence of NNs, as can be seen in the axes. As the segmentation ground truth of lesions on our dataset is not available, most tasks based on datasets with pixel-level ground truth will be turned into segmentation tasks instead of classification or prediction. It can be found in all the figures that with the increase in confidence, the number of high-intensity pixels increases. We also applied an algorithm to merge the saliency maps generated from different nodes of the neural

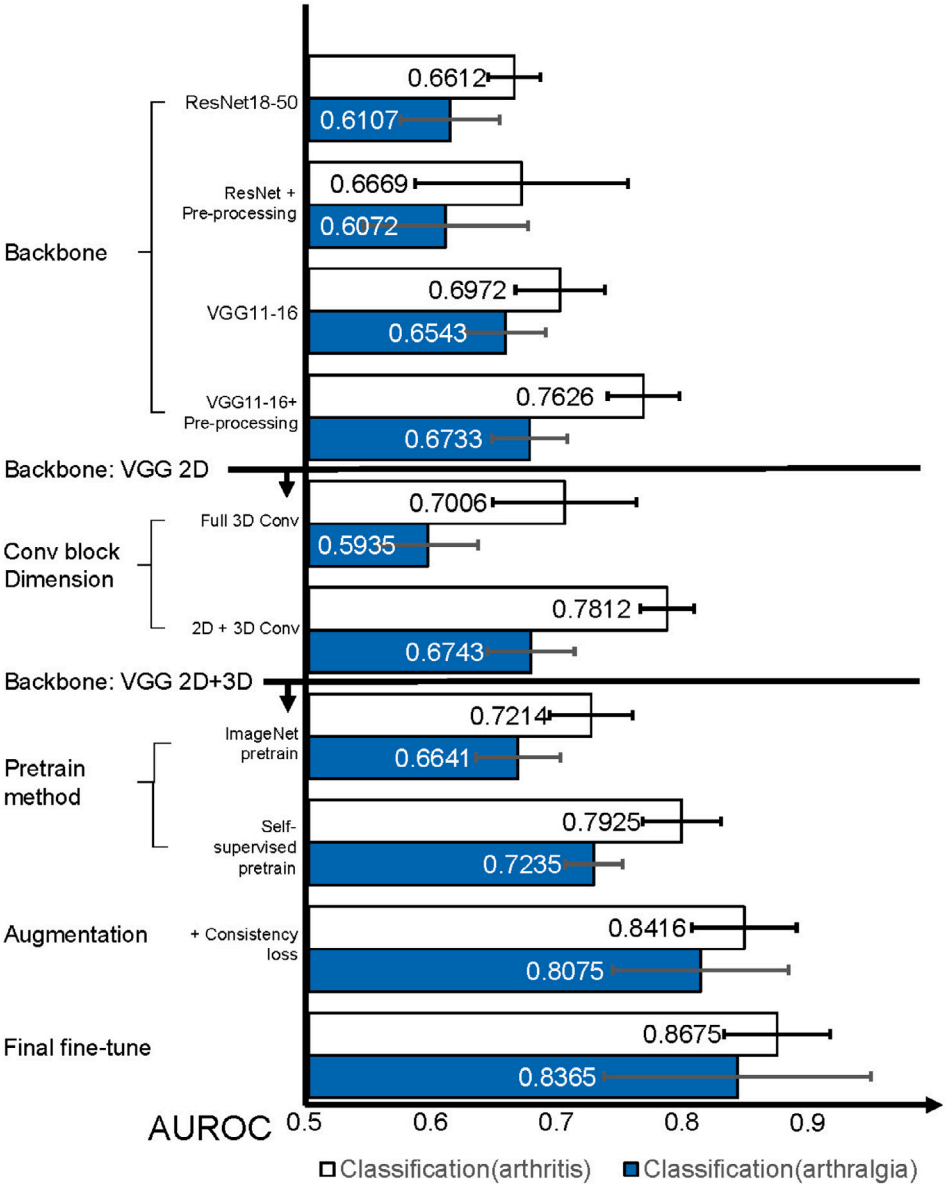

Fig. 8. Contribution of each proposed strategy.

Table 3
Best results available based on different models on each task.

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

networks and normalized them to the range of 0 and 1. Therefore, saliency maps were normalized through the max–min normalization, while the scores of the air always represent the correlation of zero. This leads to the high intensity of air caused by the normalization in some images because the scans contributed very little. That is the reason for the high values in the air in saliency maps from Task 4.

5. Discussion

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

From a clinical perspective, the DL models obtained reasonable results in distinguishing EAC, CSA and healthy controls, indicating the potential of applying DL models to assist in detecting the development of arthritis and CSA. However, the performance of DL models in distinguishing RA from other arthritides requires further investigation and improvement to assist in arthritis identification. The difference between the first two tasks and classification of RA demonstrates that the models rely on inflammatory signs. The performance of the models in RA prediction, with AUROCs of 70% on this challenging task,

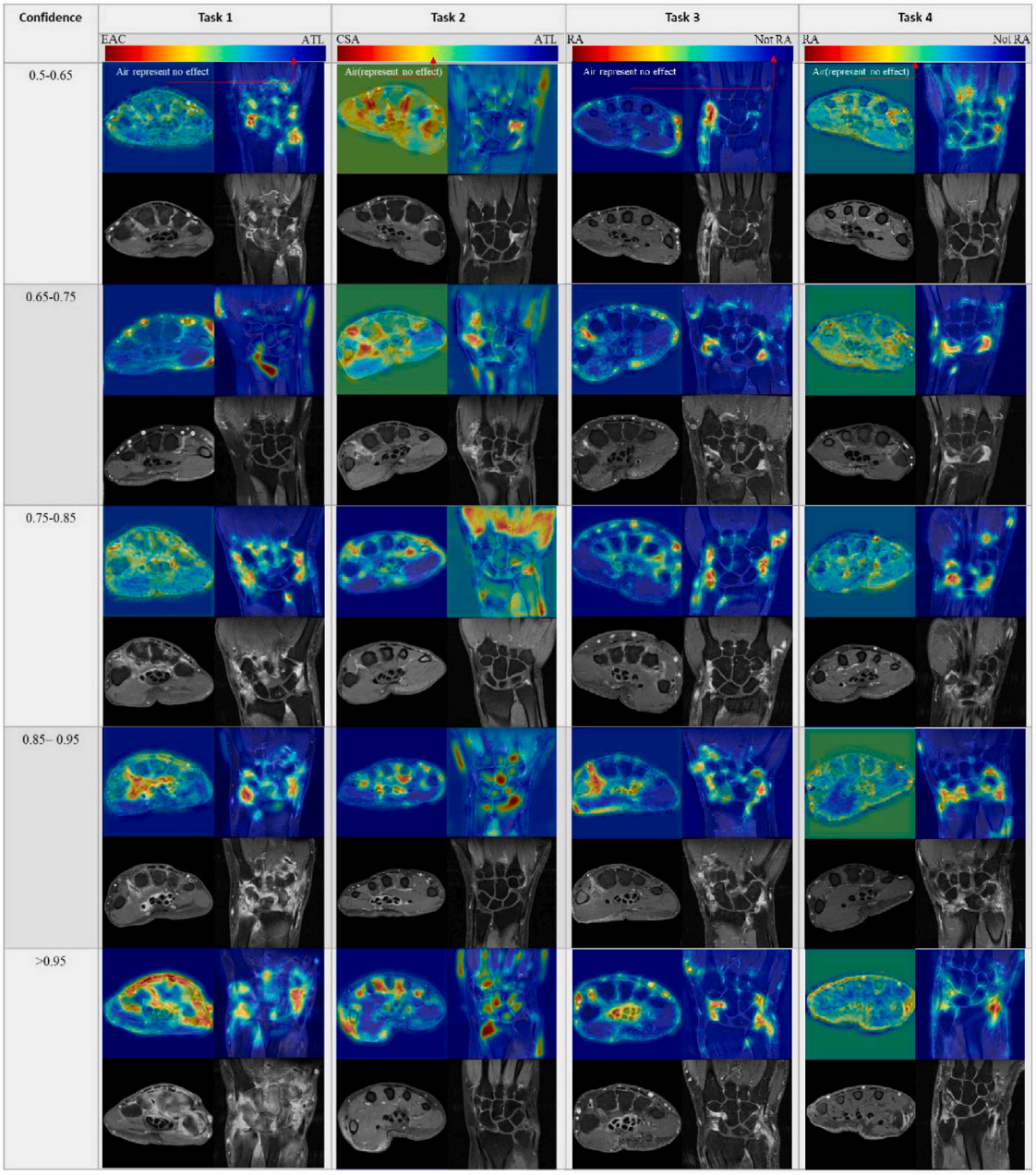

Fig. 9. Saliency maps for the four tasks. The closer the color is to red, the more the pixel contributes to the model decision for EAC, CSA or RA. Compared to the original Grad CAM, the background in the saliency maps generated in our method has a value larger than zero (yet still represents background contribution), as the threshold of no contribution to the targeted category, pixels with values less than this threshold have negative contributions, while the opposite ones have positive contributions to the model decision for this category.

is comparable to the performance by statistical analysis of clinical variables, demonstrating the potential of using DL models to search for early RA signs. According to the saliency maps, inflammatory signs appear to be the most contributing factor in current models, which is consistent with clinical knowledge. However, due to the limited dataset capacity, data imbalance and lack of other RA datasets for general validation, clinical application requires further investigation and validation.

From the perspective of method, the consistency-based strategies significantly improved the overall performance of the baseline model on all tasks (See Fig. 7). The self-supervised reconstruction, based on the consistency of specific anatomical structures, provides a pre-training method for CNNs when the amount of data is limited and when there is no similar dataset available for pre-training. As far as we know, our method is the first DL-based method for the detection of early signs of RA from MRI. Most previous studies related to RA were based on Ultrasound [37,38] or X-rays [39] and were focusing on predicting visual scores. Moreover, most other studies for MRI-based diagnosis are based on the combination of prior knowledge and the variants of standard ResNets and VGGs. These backbone models have therefore been included in our method comparison, as specific prior knowledge in these fields cannot be transferred. Compared to other methods, our methods focus more on learning from the information consistency, namely unchanged anatomical structures, and labels during processing (augmentation and DL-based feature extraction). These strategies are transferable, as they are generally not dependent on any specific prior knowledge. However, they are very dependent on the reconstruction quality and augmentation methods and therefore rely

on anatomical consistency and disease severity consistency during the spatial transformation.

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

Meanwhile, apart from labels, deep learning models are independent of other expert knowledge, we expect more information than the prediction of labels. With a verified visualization method that can test the reliability of models, the deep learning models can also serve as a way of exploring inflammatory signs of RA that have not been considered by clinical observers but could still be relevant to RA development (i.e., hypothesis-free interpretation). Combined with the current visualization methods, as shown in Fig. 9, the saliency maps can already illustrate some potential regions where some early signs of RA exist. With our ongoing studies on improving visualization methods, it has become feasible to generate saliency maps and find early signs of RA. This may give a different perspective for studying RA or finding potential image biomarkers for early RA.

6. Conclusion

As far as we know, our method is the first DL-based method for detecting the early signs of RA from MRI. Our models, based on the proposed consistency-based strategies, succeeded in all four RA classification/prediction tasks. This indicates the existence of early signs of RA in MRIs and it demonstrates the potential of DL models in RA-related research. The proposed model could serve as an initial DL benchmark in RA prediction based on MRI and indicates the ability of DL to assist RA analysis and finding early signs of RA in MRI scans with the visualization method, contributing to both technical and clinical RA studies in the future.

CRediT authorship contribution statement

Yanli Li: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Tahereh Hassanzadeh:** Conceptualization, Writing – review & editing. **Denis P. Shamonin:** Conceptualization, Software, Writing – review & editing. **Monique Reijnierse:** Supervision, Writing – review & editing. **Annette H.M. van der Helm-van Mil:** Funding acquisition, Supervision, Validation, Writing – review & editing. **Berend C. Stoel:** Conceptualization, Data curation, Funding acquisition, Investigation, Projectadministration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: AHM van der Helm-van Mil reports financial support was provided by European Research Council. Berend C. Stoel reports financial support was provided by Netherlands Organization for Scientific Research(NWO). Yanli Li reports financial support was provided by China Scholarship Council.

Data availability

The authors do not have permission to share data.

Acknowledgments

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2024.105990.

References

[1] Vikas Majithia, Stephen A. Geraci, Rheumatoid arthritis: Diagnosis and management, Am. J. Med. 120 (2007) 936–939.
[2] Michael P.M. Van Der Linden, Saskia Le Cessie, Karim Raza, Diane Van Der Woude, Rachel Knevel, Tom W.J. Huizinga, Annette H.M. Van Der Helm-Van Mil, Long-term impact of delay in assessment of patients with early arthritis, Arthritis Rheum. 62 (2010) 3537–3546.
[3] Mikkel Østergaard, Charles Peterfy, Philip Conaghan, Fiona McQueen, Paul Bird, Bo Ejbjerg, Ron Shnier, Philip O'Connor, Mette Klarlund, Paul Emery, Harry Genant, Marissa Lassere, John Edmonds, OMERACT rheumatoid arthritis magnetic resonance imaging studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system, J. Rheumatol. 30 (6) (2003) 1385–1386.
[4] Evgeni Aizenberg, Denis P. Shamonin, Monique Reijnierse, Annette H.M. van der Helm-van Mil, Berend C. Stoel, Automatic quantification of tenosynovitis on MRI of the wrist in patients with early arthritis: a feasibility study, Eur. Radiol. 29 (2019) 4477–4484.
[5] M.L. Hetland, B. Ejbjerg, K. Hørslev-Petersen, S. Jacobsen, A. Vestergaard, A.G. Jurik, K. Stengaard-Pedersen, P. Junker, T. Lottenburger, I. Hansen, L.S. Andersen, U. Tarp, H. Skjødt, J.K. Pedersen, O. Majgaard, A.J. Svendsen, T. Ellingsen, H. Lindegaard, A.F. Christensen, J. Vallø, T. Torfing, E. Narvestad, H.S. Thomsen, M. Østergaard, MRI bone oedema is the strongest predictor of subsequent radiographic progression in early rheumatoid arthritis. Results from a 2-year randomised controlled trial (CIMESTRA), Ann. Rheum. Dis. 68 (2009) 384–390.
[6] Fan Xiao, James F. Griffith, Andrea L. Hilkens, Jason C.S. Leung, Jiang Yue, Ryan K.L. Lee, David K.W. Yeung, Lai Shan Tam, ERAMRS: a new MR scoring system for early rheumatoid arthritis of the wrist, Eur. Radiol. 29 (2019) 5646–5654.
[7] Pernille Bøyesen, Espen A. Haavardsholm, Mikkel Østergaard, Désirée Van Der Heijde, Sølve Sesseng, Tore K. Kvien, Mri in early rheumatoid arthritis: Synovitis and bone marrow oedema are independent predictors of subsequent radiographic progression, Ann. Rheum. Dis. 70 (2011) 428–433.
[8] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, Olaf Ronneberger, Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature Medicine 24 (2018) 1342–1350.
[9] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.
[10] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, Dale R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA - Journal of the American Medical Association 316 (2016) 2402–2410.
[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick, Masked autoencoders are scalable vision learners, 2021.
[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[13] Evgeni Aizenberg, Edgar A.H. Roex, Wouter P. Nieuwenhuis, Lukas Mangnus, Annette H.M. van der Helm-van Mil, Monique Reijnierse, Johan L. Bloem, Boudewijn P.F. Lelieveldt, Berend C. Stoel, Automatic quantification of bone marrow edema on MRI of the wrist in patients with early arthritis: A feasibility study, Magn. Reson. Med. 79 (2018) 1127–1134.
[14] Y. Li, D. Shamonin, T. Hassanzadeh, M. Reijnierse, A. Van der Helm – van Mil, B. Stoel, Op0002 exploring the use of artificial intelligence in predicting rheumatoid arthritis, based on extremity MR scans in early arthritis and clinically suspect arthralgia patients, Ann. Rheum. Dis. 82 (Suppl 1) (2023) 1–2.
[15] P. Maragos, Differential morphology and image processing, IEEE Trans. Image Process. 5 (1996) 922–937.
[16] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, Klaus H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021) 203–211.
[17] M. Firouzi, S. Fadaei, A. Rashno, A new framework for Canny edge detector in hexagonal lattice, Int. J. Eng. 35 (8) (2022) 1588–1598.
[18] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015.
[19] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, Jianming Liang, Models genesis, 2020.
[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A simple framework for contrastive learning of visual representations, 2020.
[21] Aaron van den Oord, Yazhe Li, Oriol Vinyals, Representation learning with contrastive predictive coding, 2018.
[22] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet: A unified embedding for face recognition and clustering, 2015.
[23] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, Vineeth N Balasubramanian, Grad-CAM++: Improved visual explanations for deep convolutional networks, 2017.
[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, 2016.
[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization, 2015.
[26] Xanthe ME Matthijssen, Fenne Wouters, Debbie M Boeters, Aleid C Boer, Yousra J Dakkak, Ellis Niemantsverdriet, Annette HM van der Helm-van Mil, A search to the target tissue in which RA-specific inflammation starts: a detailed MRI study to improve identification of RA-specific features in the phase of clinically suspect arthralgia, Arthritis Res. Therapy 21 (2019) 1–11.
[27] Xue Zhao, Jing-Wen Bai, Qiu Guo, Ke Ren, Guo-Jun Zhang, Clinical applications of deep learning in breast MRI, Biochim. Biophys. Acta (BBA) - Rev. Cancer 1878 (2) (2023) 188864.
[28] Shakila Shojaei, Mohammad Saniee Abadeh, Zahra Momeni, An evolutionary explainable deep learning approach for Alzheimer's MRI classification, Expert Syst. Appl. 220 (2023) 119709.
[29] Mubashir Hussain, Deepika Koundal, Jatinder Manhas, Deep learning-based diagnosis of disc degenerative diseases using MRI: A comprehensive review, Comput. Electr. Eng. 105 (2023) 108524.
[30] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, CoRR, 2017, abs/1709.01507.
[31] Lukas Folle, Sara Bayat, Arnd Kleyer, Filippo Fagni, Lorenz A Kapsner, Maja Schlereth, Timo Meinderink, Katharina Breininger, Koray Tascilar, Gerhard Krönke, Michael Uder, Michael Sticherling, Sebastian Bickelhaupt, Georg Schett, Andreas Maier, Frank Roemer, David Simon, Advanced neural networks for classification of MRI in psoriatic arthritis, seronegative, and seropositive rheumatoid arthritis, Rheumatology (2022).
[32] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, CoRR, 2017, abs/1704.04861.
[33] Sachin Mehta, Mohammad Rastegari, MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, CoRR, 2021, abs/2110.02178.
[34] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, Pinkie Anggia, Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer, Procedia Comput. Sci. 179 (2021) 423–431, 5th International Conference on Computer Science and Computational Intelligence 2020.
[35] Sarmad Maqsood, Robertas Damasevicius, Faisal Mehmood Shah, An efficient approach for the detection of brain tumor using fuzzy logic and U-NET CNN classification, in: Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part V 21, Springer, 2021, pp. 105–118.
[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, CoRR, 2020, abs/2010.11929.
[37] R.J. Hemalatha, V. Vijaybaskar, T.R. Thamizhvani, Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning, Proc. Inst. Mech. Eng. H J. Eng. Med. 233 (6) (2019) 657–667.
[38] Jakob Kristian Holm Andersen, Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Kathrine Holtz, Jakob Grauslund, Thiusius Rajeeth Savarimuthu, Søren Andreas Just, Neural networks for automatic scoring of arthritis disease activity on ultrasound images, RMD Open 5 (1) (2019) e000891.
[39] George P. Avramidis, Maria P. Avramidou, George A. Papakostas, Rheumatoid arthritis diagnosis: Deep learning vs. humane, Appl. Sci. 12 (1) (2022) 10.