Computerized Medical Imaging and Graphics 107 (2023) 102239

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

A clinically motivated self-supervised approach for content-based image retrieval of CT liver images

Kristoffer Knutsen Wickstrøm [a,*], Eirik Agnalt Østmo [a], Keyur Radiya [b], Karl Øyvind Mikalsen [a,b], Michael Christian Kampffmeyer [a,c], Robert Jenssen [a,c,d]

[a] Machine Learning Group at the Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø NO-9037, Norway
[b] Department of Gastrointestinal Surgery, University Hospital of North Norway (UNN), Tromsø, Norway
[c] Norwegian Computing Center, Department SAMBA, P.O. Box 114 Blindern, Oslo NO-0314, Norway
[d] Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 København Ø, Denmark

ARTICLE INFO

Keywords:
Content-based image retrieval
Self-supervised learning
CT liver imaging
Explainability

ABSTRACT

Deep learning-based approaches for content-based image retrieval (CBIR) of computed tomography (CT) liver images is an active field of research, but suffer from some critical limitations. First, they are heavily reliant on labeled data, which can be challenging and costly to acquire. Second, they lack transparency and explainability, which limits the trustworthiness of deep CBIR systems. We address these limitations by: (1) Proposing a self-supervised learning framework that incorporates domain-knowledge into the training procedure, and, (2) by providing the first representation learning explainability analysis in the context of CBIR of CT liver images. Results demonstrate improved performance compared to the standard self-supervised approach across several metrics, as well as improved generalization across datasets. Further, we conduct the first representation learning explainability analysis in the context of CBIR, which reveals new insights into the feature extraction process. Lastly, we perform a case study with cross-examination CBIR that demonstrates the usability of our proposed framework. We believe that our proposed framework could play a vital role in creating trustworthy deep CBIR systems that can successfully take advantage of unlabeled data.

1. Introduction

Content-based image retrieval (CBIR) is a core research area in medical image analysis, with numerous studies across many different image modalities (Barata and Santiago, 2021; Ramalhinho et al., 2021; Haq et al., 2021). CBIR supports clinicians in retrieving relevant images from a large database compared to a query image, which reduces labor-intensive manual search and aids in diagnosis. For instance, a physician might want to investigate how patients in a large database with a similar disease as a new patient, such as liver metastasis, were diagnosed. The information from the previous diagnoses can then be used to determine the proper treatment for the new patient. In analysis of computed tomography (CT) images of the liver, CBIR have been an active and important area of medical image analysis for many years (Zhao et al., 2004; Chi et al., 2013; Yoshinobu et al., 2020). CBIR has the potential to make labor intensive tasks in the clinical workflow more time efficient, as illustrated in Section 8.3. Automatic support systems such as CBIR is becoming increasingly important in clinical liver disease diagnosis (Radiya et al., 2023), which is discussed in detail in Section 2.1.

Currently, deep learning-based CBIR, or deep CBIR, constitute the state-of-the-art for CBIR (Silva et al., 2020; Yoshinobu et al., 2020; Haq et al., 2021), due to its high precision and efficiency. However, deep CBIR suffers from some critical limitations. First (1), current deep CBIR for CT liver images rely on labeled data for training (Yoshinobu et al., 2020). Obtaining labeled data can be costly and time-consuming, which therefore limits the usability of deep CBIR systems. However, recent works have shown how self-supervised learning can leverage unlabeled data to improved CBIR systems (Siradjuddin et al., 2019; Monowar et al., 2022), but such approaches have not been explored in the context of CBIR of CT liver images. Second (2), deep CBIR suffer from a fundamental lack of explainability. This can have detrimental effects in a clinical setting, since deep learning-based systems are known to exploit confounding factors and artifacts to make their predictions. For instance, Gautam et al. (2022) showed that a deep-learning-based system learned to use tokens and artifacts in X-ray images to make its predictions instead of clinically relevant features. These tokens and artifacts would not be present for new patients, and such a system would not work as intended if put into clinical practice. Therefore, it is

* Corresponding author.
E-mail address: kristoffer.k.wickstrom@uit.no (K.K. Wickstrøm).

https://doi.org/10.1016/j.compmedimag.2023.102239
Received 6 June 2022; Received in revised form 2 May 2023; Accepted 2 May 2023
Available online 9 May 2023

not advisable to blindly trust the retrieved images from the deep CBIR system without investigating what input features influence the retrieval process through an explainability analysis.

A promising direction to address the first limitation is learning from unlabeled data through self-supervision. Recent self-supervised learning frameworks have shown remarkable results, in some cases even rivaling supervised learning (Chen et al., 2020; Caron et al., 2020; Chen and He, 2021). In a nutshell, contemporary self-supervised approaches train a feature extractor that extracts informative representations by exploiting known invariances in the data. These representations can then be used for other tasks, such as CBIR by taking similarities between the new representations to retrieve similar images. These self-supervised approaches have been show to improve performance in the context of chest X-ray (Truong et al., 2021; Azizi et al., 2021), dermatology classification (Azizi et al., 2021), organ and cardiac segmentation (Hansen et al., 2022), and whole heart segmentation (Dong et al., 2021), but have yet to be developed for CBIR of CT liver images.

In this paper, we propose a clinically motivated self-supervised framework for CBIR of CT liver images. Our proposed framework incorporates domain knowledge that exploits known properties of the liver, which leads to improved performance compared to well-known self-supervised baselines. Concretely, a novel Hounsfield unit clipping strategy that removes non-liver pixels from the input and allows the system to focus on the liver is incorporated into the self-supervised training. While the focus in this paper is on the liver in CT images, our proposed framework could also be used to focus on other organs by altering how the Hounsfield units are clipped. For the second limitation, great improvements have been made in the field of explainable artificial intelligence (XAI) over the last couple of years, and numerous studies have shown how XAI can improve the reliability and trustworthiness of deep learning-based systems in healthcare (Silva et al., 2020; Gautam et al., 2022; Wickstrøm et al., 2021; Sayres et al., 2019). However, the majority of these improvements have been in algorithms that can explain models which produce decisions, such as classification or similarity scores. When learning from unlabeled data through e.g. self-supervised learning, such a score or similarity measure might not be available and standard XAI techniques cannot be applied. But the recent field of representation learning explainability (Wickstrøm et al., 2023) aims at explaining vector representations, and can therefore tackle the lack of explainability in deep CBIR. But such a representation learning explainability analysis has not been performed in the context of CBIR of CT liver images.

Our contributions are:

- A clinically motivated self-supervised framework specifically designed to extract features focused on the liver.
- A novel analysis that explains the representations produced in the feature extraction process.
- Thorough evaluation on real-world datasets.
- A case-study where images from the same patient are retrieved across different examinations.

2. Related work

2.1. Clinical background for automatic support systems in liver disease research

Performing a liver biopsy is the gold standard for assessing the nature and severity of liver disease (Bravo et al., 2001), but it also carries the risk of complications during the procedure (Tapper and Lok, 2017). Noninvasive evaluation offers an alternative to performing a biopsy that avoids this risk for complications, where CT imaging have been a popular noninvasive approach to conduct liver disease diagnosis (Tapper and Lok, 2017). Manual evaluation of CT images by clinicians is common in clinical practice, but is labor intensive and for challenging for particular types of disease (Tapper and Lok, 2017). Therefore, it is desirable to design automatic systems that could support and assist clinicians in noninvasive evaluation of liver disease. Automatic systems have demonstrated great potential in supporting clinicians for several important liver diseases. Below we discuss recent advances and important aspects of automatic support systems in liver disease research.

Focal liver lesions (FLLs) is a common occurrence in clinical practice. To distinguish between FLLs is crucial in daily clinical practice to address the treatment accordingly. Incidence of FLLs was encountered in around 28.5% (Kreft, 2001) for less than or equal to 2 cm in diameter. The ratio would be higher when including the lesions above 2 cm. Diagnosis of FLLs is cumbersome, especially in the absence of malignancy in the patient's history. Though incidental malignant lesions are not uncommon (Kreft, 2001), CT images play a significant role in diagnosing FLLs, while percutaneous biopsy in cases of doubt is considered the gold standard method. However, fine needle aspiration biopsy and cut needle biopsy have an accuracy of 78% in diagnosing FLL (França et al., 2003). Major and minor complications related to liver biopsy have been identified at 2.44% and 9.53%, consecutively and remains constants in last decades (Thomaides-Brears et al., 2021). However, biopsies are relatively contraindicated in certain types of cancer due to the risk of tumor seeding or fatal bleeding from FLLs. Noninvasive methods have shown comparable results to biopsy for diagnosing liver diseases, thus becoming preferable methods for diagnosing certain types of lesions (Tapper and Lok, 2017). A recent preliminary study by Tiyarattanachai et al. (2022) demonstrated that an automatic FLL detection system could result in similar performance as radiologists, but operation speed reached 30–34 frames per second. CBIR facilitates the physician or radiologist to identify the cases diagnosed and treated for similar FLLs for case studies. CBIR could also play a role in creating an image database of similar lesions for the development of DL-based diagnostic tools.

Chronic liver disease (CLD) is a progressive disease that gradually deteriorates the liver function, and can be challenging to identify at an early stage (Zheng et al., 2022). Automatic detection systems could aid in early identification and treatment of CLD. Singal et al. (2013) found that machine learning systems outperformed traditional modeling methods in predicting development of hepatocellular carcinoma. Diffuse liver disease staging is another important area of liver disease research where automatic support systems have been developed (Zhou et al., 2019). Yasaka et al. (2018) investigated the performance of deep learning in the staging of liver fibrosis, and found that the automatic system exhibited a high diagnostic performance. Wang et al. (2018) conducted a prospective to evaluate the performance of deep learning for assessing liver fibrosis stages and found that the system could provide good overall performance, and that diagnostic accuracy improves as more image were acquired from each individual.

The lack of explainability is regularly highlighted as a major obstacle for the effective implementation of automatic support systems in the healthcare sector (Marwaha and Kvedar, 2022; Chen et al., 2022; Ching et al., 2018). This has also been highlighted in the context of liver disease research (Nam et al., 2022), but no systems have been developed and tested through clinical trials in the context of CT liver images. Studies have shown that automatic systems that provide an explanation together with their prediction can improve performance, for instance in grading of diabetic retinopathy (Sayres et al., 2019). Therefore, it is essential to include explainability in the development of automatic support systems.

2.2. Content-based image retrieval

The goal of CBIR is to find similar images from a large-scale database, given a query image. CBIR is an active area of research that span numerous medical imaging domains, such as X-ray (Haq et al., 2021; Silva et al., 2020), dermatology (Barata and Santiago, 2021; Ballerini et al., 2010), mammography (Jiang et al., 2014), and histopathology (Peng et al., 2019; Zheng et al., 2019). An illustration of a CBIR system in the context of CT liver images is shown in Fig. 1.

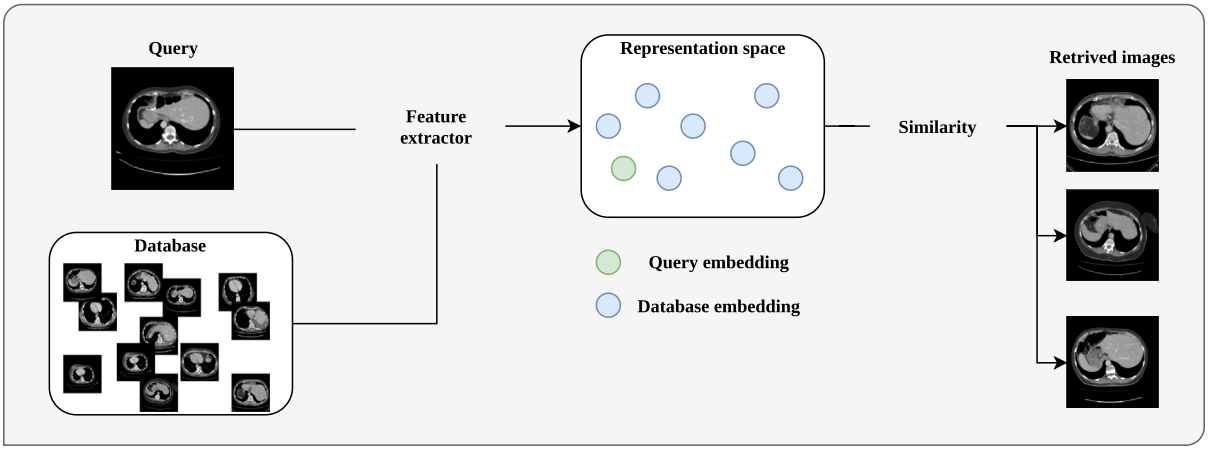

Fig. 1. Illustration of content-based image retrieval.

2.3. Content-based image retrieval of CT liver images

CBIR of CT liver images have been extensively studied. Early studies relied on handcrafted features based on certain properties in the images. Gabor filters have been used to extract texture information (Zhao et al., 2004). Texture information have also been combined with density information in the context of focal liver lesion retrieval (Chi et al., 2013). Manifold learning have been utilized to facilitate CBIR of CT liver images (Mirasadi and Foruzan, 2019). Lastly, a Bayesian approach has been studied in connection with multi-labeled CBIR of CT liver images (Ramalhinho et al., 2021).

Recently, deep learning-based feature extraction have improved performance significantly in CBIR of CT liver images. The most straight forward approach for deep CBIR is to train a neural network for the task of CT liver image classification and use the intermediate features prior to the classification layer for calculating similarities. This has been demonstrated to produce good results when the network was trained for the task of focal liver lesions detection (Yoshinobu et al., 2020). However, all these approaches need labeled data to train the deep learning-based feature extractor.

2.4. Self-supervised learning

Learning from unlabeled data is a fundamental problem in machine learning. Recently, self-supervised learning have shown promising results in computer vision (Chen et al., 2020; Chen and He, 2021), natural language processing (Devlin et al., 2019; Brown et al., 2020), and time series analysis (Franceschi et al., 2019; Wickstrøm et al., 2022). Furthermore, recent studies have also demonstrated that self-supervised learning can improve performance across several imaging domains in medical image analysis (Azizi et al., 2021; Truong et al., 2021; Hansen et al., 2022; Dong et al., 2021).

For computer vision, there are three main approaches to self-supervised learning. First, contrastive self-supervised learning is performed by sampling positive pairs and negative samples and learning a representation where the positive pairs are mapped in close proximity and far from the negative samples. The SimCLR framework (Chen et al., 2020) is one of the most widely used approaches in this category. Second, clustering-based self-supervised learning utilizes clustering algorithms to produce pseudo-labels which in turn are used to learn a useful representation of the data. DeepCluster (Caron et al., 2018) and the SwAV framework (Caron et al., 2020) are two of the most widely used clustering-based self-supervised approaches in the literature. Lastly, siamese self-supervised approaches learn how to produce a useful representation by maximizing agreement between positive pairs of samples. The two main contemporary approaches in siamese self-supervised approaches are the SimSiam framework (Chen and He, 2021) and the BYOL framework (Grill et al., 2020).

2.5. Explainability

Explainability is of vital importance for machine learning systems in healthcare. Without it, clinicians cannot fully trust the algorithms decision and the system becomes less reliable. Many recent studies have shown how explainability can be incorporated into deep learning systems for medical image analysis, ranging from diabetic retinopathy (Quellec et al., 2021), dermatology (Barata and Santiago, 2021; Gu et al., 2021), X-ray (Khakzar et al., 2021), and endoscopic images (Wickstrøm et al., 2020; Vasilakakis et al., 2021).

Most of the widely used explainability techniques typically leverage the classification or similarity score to ascertain input feature importance (Springenberg et al., 2015; Schulz et al., 2020; Plummer et al., 2020), and such approaches have been explored in the context of deep CBIR. For models trained for classification tasks, explanations through gradient information have been shown to both provide new insights and improve performance for X-ray images (Silva et al., 2020). For models trained to output a similarity score, it has been shown how the similarity score can be used to provide explanations (Dong et al., 2019; Plummer et al., 2020). Similarity score explanations have been explored for X-ray images (Hu et al., 2022). Lastly, it has been shown that explanation by examples can be effective in histopathological images (Peng et al., 2019).

In the unlabeled setting where only the feature extraction model is available, these techniques are not applicable. In such cases, it is desirable to explain the vector representation of an image, since the decision is not available. Representation learning explainability is a very recent field of XAI, that has yet to be developed for CBIR. In this work, we leverage the RELAX framework (Wickstrøm et al., 2023) to explain the feature extractors trained using self-supervised learning. RELAX is the first method that allows for representation learning explainability and has been shown to provide superior performance to competing alternatives (Wickstrøm et al., 2023).

3. A clinically motivated self-supervised approach for CT liver images

In this section, we present our proposed clinically-motivated self-supervised approach and the SimSiam framework for self-supervised learning.

3.1. A clinically motivated self-supervised approach for CT liver images

We propose to incorporate clinical knowledge into our self-supervised framework to learn more clinically relevant features. In self-supervised learning, known invariances in the data are used to train a feature extractor that extracts relevant features from the input

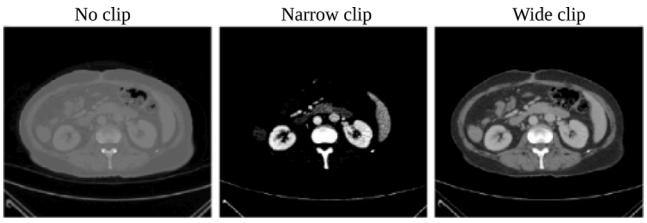

Fig. 2. Effect of Hounsfield unit clipping on CT liver images. From left to right, no clipping, narrow clip (50, 150), and wide clip (−200, 300).

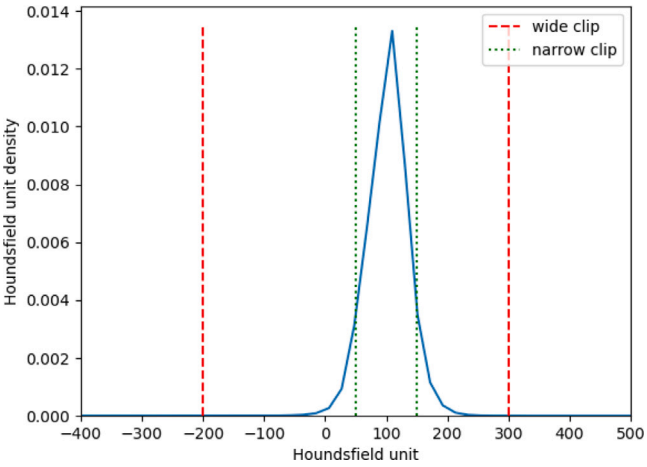

Fig. 3. Distribution of pixel intensity values for liver pixel from the Decathlon dataset and the two clipping strategies used in our proposed framework.

images. For instance, the liver can occur on both the left and right hand side of an image, depending on which direction the patient is inspected. Therefore, the feature extractor should be invariant to horizontal flips in the images, and this invariance can be learned by incorporating horizontal flipping into the self-supervised learning procedure. Identifying these invariances is crucial to make the self-supervised system work properly and focus on clinically relevant features in the input images. Our motivation is based on the knowledge that the pixel intensities of the liver lay within a certain range for CT images. A standard pre-processing step is to clip the pixel intensities of the CT images (Li et al., 2018a), such that unimportant pixels are removed prior to learning. The pixel intensities of CT images represent a physical quantity, namely the Hounsfield unit. The same clipping is usually applied to all images. However, if this clipping was incorporated into the self-supervised learning procedure, the network could be guided to learn which features are liver features and which ones are not. Our motivation is to exploit the knowledge that the liver should be invariant to pixel intensity clipping for a certain range of clipping.

Based on this motivation, we propose a Hounsfield clipping strategy where the pixel values for the same image are clipped and scaled based on different ranges of Hounsfield units. Fig. 2 shows how our proposed clipping scheme affects an image. The leftmost image has no clipping applied, and illustrates why it is important to remove some pixel intensities in order to highlight relevant structures in the images. The middle image shows the narrow clipping strategy between 50 and 150 Hounsfield units. Notice how only the liver and some other organs are now visible in the image. The rightmost image shows the wide clipping strategy between −200 and 300 Hounsfield units. In this case, some redundant structures are removed, but more organs are left visible compared to the middle image. The images considered in this paper are intra venous contrast enhanced images taken in the portal venous phase. These two ranges were chosen based on the following. First, it is known that the liver typically has Hounsfield units in the range 50–60 (Tisch et al., 2019). Furthermore, we have collected all pixel intensities for the liver in the Decathlon dataset. These values are shown in Fig. 3, and illustrates how the narrow clip will remove some of the liver pixels but keep the main proportion, while the wide clip will keep almost all liver pixels apart from some outliers. Our proposed framework for learning representations that focus on liver features is shown in Fig. 4. Each image is clipped with the wide and narrow range, before the data augmentation is applied. Afterwards, we follow the SimSiam approach described below. During testing, we use the wide clipping to ensure that most liver pixels are kept in the images. Our proposed clipping procedure introduced minimal computational overhead, due to the clipping operation being simple to compute. Also, since the same architecture is used for the feature extraction, our proposed framework has the same computational demand and number of parameters as previous self-supervised approaches.

3.2. SimSiam framework

In this work, we build on the SimSiam framework. The main motivation for this choice is that both contrastive and clustering-based self-supervised approaches requires a large batch size during training to provide high quality representations (Chen et al., 2020; Caron et al., 2020). This can be computationally challenging, especially if the medical images in question are large. However, the siamese-based approaches (Chen and He, 2021) are less sensitive to the batch size used during training. Furthermore, we opt for the SimSiam approach over BYOL to avoid training both a student and a teacher network used in BYOL, again to avoid additional computations.

Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ represent an input image with height H and width W and f a feature extractor that transforms $\mathbf{X}$ into a new d-dimensional representation $\mathbf{h} \in \mathbb{R}^d$, that is $f(\mathbf{X}) = \mathbf{h}$. Next, two views $\mathbf{X}_1$ and $\mathbf{X}_2$ are constructed by augmenting the original image. The task performed in SimSiam to learn a useful representation, is to maximize the similarity between the two views. The representation $\mathbf{h}$ is the new representation that can be used for downstream tasks, such as CBIR. However, the loss is not computed directly on the output of the feature extractor f. Instead, a multilayer perceptron-based projection head g transforms $\mathbf{h}$ into a new representation $\mathbf{z}$, that is $g(\mathbf{h}) = \mathbf{z}$, where the loss is computed. This projector is a crucial component in most self-supervised frameworks (Chen et al., 2020; He et al., 2020), as it avoids dimensional collapse in the representation $\mathbf{h}$ (Jing et al., 2022), which is the one that will be used for downstream tasks such as CBIR. The learning is performed by minimizing the negative cosine similarity between the two views:

$$D(\mathbf{z}_1, \mathbf{z}_2) = -\frac{\mathbf{z}_1}{\|\mathbf{z}_1\|_2} \cdot \frac{\mathbf{h}_2}{\|\mathbf{h}_2\|_2}, \tag{1}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. The full SimSiam loss function is defined as:

$$L = D(\mathbf{z}_1, \mathbf{h}_2) + D(\mathbf{z}_2, \mathbf{h}_1). \tag{2}$$

An important component of the SimSiam framework is a stop-gradient (stopgrad) operation, which is incorporated into Eq. (2) as follows:

$$L = \frac{1}{2} D(\mathbf{z}_1, \text{stopgrad}(\mathbf{h}_2)) + \frac{1}{2} D(\mathbf{z}_2, \text{stopgrad}(\mathbf{h}_1)) \tag{3}$$

The stop-gradient operation is applied to the projector network, such that the encoder on $\mathbf{X}_2$ no gradient from $\mathbf{h}_2$ in the first term, but it receives gradients from $\mathbf{z}_2$ (and similarly for $\mathbf{X}_1$). The stop-grad operation allows SimSiam to mimic a teacher–student setup, but avoids the need to store two networks. Furthermore, it has been shown that the stop-grad operation is critical to avoid the problem of complete collapse in the representations (Tian et al., 2021).

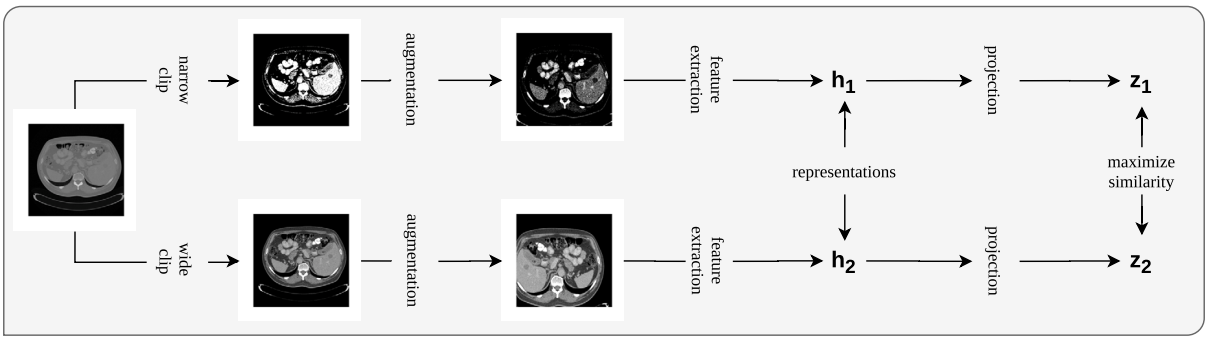

Fig. 4. Illustration of proposed self-supervised framework.

Data augmentation. The prior knowledge injected through the data augmentation is of paramount importance to ensure that the models learns relevant features. The data augmentation used in SimSiam is similar to the standard approach in recent self-supervised learning (He et al., 2020; Chen et al., 2020):

1. Crop with a random proportion from [0.2, 1.0], and resize to a fixed size.
2. Flip horizontally with a probability of 0.5.
3. Color augmentation is performed by randomly adjusting the brightness, contrast, saturation, and hue of each image with a strength of [0.4, 0.4, 0.4, 0.1]
4. Randomly convert image to gray scale version with a probability 0.2.

Note that the input images are converted to pseudo RBG images by stacking the input image 3 times along the channel axis. Prior works have shown that the augmentation scheme listed above can lead to increased performance across several medical image related tasks (Azizi et al., 2021; Truong et al., 2021; Hansen et al., 2022; Dong et al., 2021), albeit not in the context of CBIR of CT liver images. However, these augmentations are selected with natural images in mind, and do not take into account the properties of CT liver images. Our proposed Hounsfield unit clipping scheme takes into account the particular characteristics of CT images of the liver, which we hypothesize can improve the self-supervised framework.

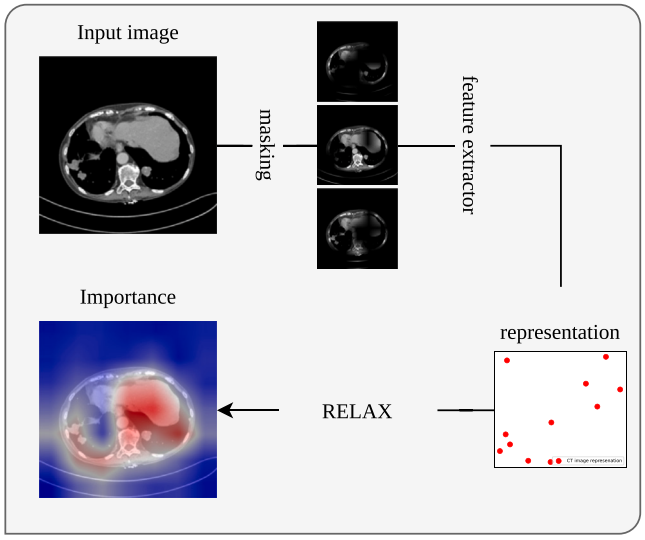

Fig. 5. Illustration of RELAX. A feature extractor produces a new representation of an input image, and RELAX determines what input features are important for the representation.

4. Explaining representations

Explainability is a critical component for creating trustworthy and reliable deep learning-based systems. For deep CBIR, we want to know what information the feature extractor is using to create the representation that the retrieval is based on. This requires explaining the vector representations produced by the feature extractor, which cannot be accomplished with standard explainability techniques since they require a classification or similarity score to create the explanation. However, the recent field of representation learning explainability address the problem of explaining representations (Wickstrøm et al., 2023). In this work, we leverage the RELAX (Wickstrøm et al., 2023) framework to explain the representations used in the CBIR system.

4.1. RELAX

RELAX is an occlusion-based explainability framework that provides input feature importance in relation to a vector representation, as opposed to a classification or similarity score. The core idea of RELAX is to evaluate how the representation of an image changes as parts of the image are removed using a mask. Let $\mathbf{M} \in [0,1]^{H \times W}$ represent a stochastic mask used for removing parts of the image. Next, $\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M})$, where $\odot$ denotes element-wise multiplication, is the representation of a masked version of $\mathbf{X}$ and $s(\mathbf{h}, \bar{\mathbf{h}})$ is a similarity measure between the unmasked and the masked representation. The intuition behind RELAX is that when informative parts are masked out, the similarity between the two representations should be low, and vice versa for non-informative parts. Finally, the importance R_{ij} of pixel (i, j) is defined as:

$$\bar{R}_{ij} = \frac{1}{N} \sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n). \tag{4}$$

Here, $\bar{\mathbf{h}}_n$ is the representation of the image masked with mask n, and $M_{ij}(n)$ the value of element (i, j) for mask n. The similarity measure used in the cosine similarity, as proposed in prior works (Wickstrøm et al., 2023). The RELAX framework is illustrated in Fig. 5.

The mask generation is a crucial component in RELAX. In this work, we follow the strategy used in previous studies (Petsiuk et al., 2018; Wickstrøm et al., 2023). Binary masks of size $h \times w$, where $h < H$ and $w < W$, are generated, where each element of the mask is sampled from a Bernoulli distribution with probability p. To produce smooth and spatially coherent masks, the small masks are upsampled using bilinear interpolation to the same size as the input image. Furthermore, the number of masks required to obtain reliable estimates of importance is an important hyperparameter. In this work, we generate 3000 masks to obtain an explanation for a single image, as suggested in a prior work (Wickstrøm et al., 2023).

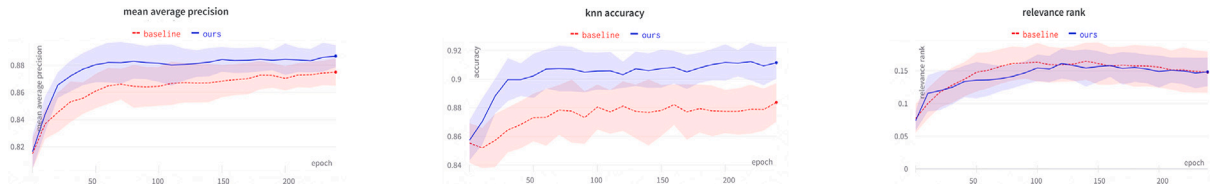

Fig. 6. From left to right, mean average precision, knn accuracy, and relevance rank scores versus epochs across 20 training runs on the test images from the Decathlon dataset The plot show how performance increase with training time, and that the proposed framework learns faster with better results.

5. Evaluation

We introduce the set of scores utilized to provided quantitative evaluation of our proposed framework.

5.1. Evaluating quality of CBIR

A standard approach to evaluate the quality of a CBIR system is to measure the class-consistency in the top retrieved images (Silva et al., 2020; Li et al., 2018b). One of the most common approaches to evaluate the class-consistency is through mean average precision (MAP):

$$\text{MAP} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\text{precision(k)}_n, \tag{5}$$

where N is the number of test samples (query images), K is the top-K retrieved images for each query image, and precision is defined as:

$$\text{precision(k)} = \frac{|\text{relevant images} \cap \text{k-retrieved images}|}{|\text{k-retrieved images}|}. \tag{6}$$

MAP evaluates the precision of the retrieved images across several values of K, which makes it robust towards fluctuations among the top retrieved images.

5.2. Evaluating quality of representations

The most widespread approach for evaluating the representation produced by a self-supervised learning framework is to train a simple classifier on the learned representations (Chen et al., 2020; Caron et al., 2020; He et al., 2020). The motivation for this, is that a simple classifier is highly dependent on the representation it is given to perform the desired task. In this work, we follow recent studies that use a k-nearest neighbors (KNN) classifier (Caron et al., 2021, 2020) to evaluate the quality of the representation. We opt for a KNN classifier over a linear classifier as it does not require any training, which can lead to ambiguities in the results (Kolesnikov et al., 2019), and has minimal hyperparameters to tune.

5.3. Evaluating the quality of explanations

Great improvements have been made in the field of XAI over the last couple of years. In contrast, the field of evaluation for explanations is still under active development (Doshi-Velez and Kim, 2017). However, recent advances have introduced new methods for providing quantitative evaluation of explanations. In this work, we use the relevance rank accuracy score (RR) (Arras et al., 2022). RR measures how many of the top-M relevant pixels lies within the ground truth segmentation mask. It can be considered a proxy for how well the explanation agrees with a human explanation for a given images. Let R_M denote the M most relevant pixels in an explanation, and S the segmentation mask for the liver. RR can then be defined as:

$$\text{RR} = \frac{1}{N}\sum_{n}^{N}\frac{|R_M(n) \cap S(n)|}{|S(n)|}. \tag{7}$$

The RR is computed using the Quantus toolbox (Hedström et al., 2023).

5.4. Statistical test for quantitative evaluation

We use a permutation test (Welch, 1990) to determine if differences in quantitative performance is statistically significant. The null hypothesis in a permutation test is that all samples come from the same distribution. In our case, the null hypothesis is that the performance of two feature extractors trained using different methods is the same, and the test statistic is the difference between mean performance across multiple training runs. If the performance is truly equal, the mean difference should not change significantly if the performance of the two feature extractors are mixed together and shuffled. The p-value is computed by counting the number of mean differences that have a greater value than the mean difference of the original configuration. We repeat the training procedure 20 times for each approach, which would results in 40! permutations. This is computationally intractable, so we perform 10000 permutations of the performance scores to conduct the statistical test.

6. Data

In this section, we present the data used to evaluated our proposed framework.

6.1. Decathlon data

The medical segmentation decathlon is a biomedical image analysis challenge where several tasks and modalities are considered (Antonelli et al., 2022). One of the datasets in this challenge is a CT liver dataset acquired from the LiTS dataset (Bilic et al., 2023) and consists of 201 contrast-enhanced CT liver images from patients with mostly cancers and metastatic liver disease. However, we exclude 70 of these images as they do not include label information. Using every slide from each volume is computationally intractable. Therefore, we construct a slice-wise dataset as follows. From each volume, we sample 5 slices with no liver and 5 slices with liver. We construct the training set from the first 100 volumes and the test set from the remaining 31 volumes. This results in a balanced dataset with 1000 training images and 310 test images.

6.2. UNN data

The UNN dataset is from an extensive database of CT scans from The University Hospital of North Norway (UNN). It is under development through a close collaboration between UiT, The Arctic University of Norway, and UNN. The database contains CT volumes of 376 patients surgically treated for rectal cancer from 2006 to 2011 in North Norway. The examinations were conducted for diagnostic and routine follow-up purposes. The full dataset consists of CT with coronal, sagittal, and axial slices of mainly the thorax, abdomen, and pelvis. Examinations were conducted with different scanners and protocols at eight different hospitals in North Norway in the period 2005 to 2020.

From the full dataset a subset of 3347 axial volumes from 368 patients was selected based on descriptive keywords and DICOM metadata to limit it to contain mostly volumes of the liver and abdomen. This subset is similar to the CT liver partition of the medical segmentation decathlon dataset in terms of image resolutions and contents, but more

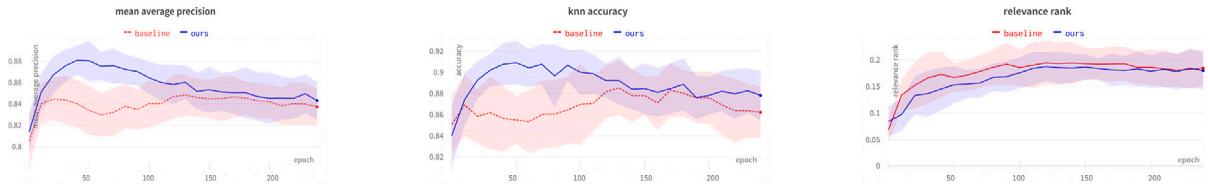

Fig. 7. From left to right, mean average precision, knn accuracy, and relevance rank scores versus epochs across 20 training runs on the test images from the UNN dataset. The plot show how performance increase with training time, and that the proposed framework learns faster with better results.

Table 1
An overview of deep learning architecture, number of parameters, activation function, and optimization method used in the proposed method.

Architecture	Num. of params.	Activation	Optimization
ResNet50	~23 million	ReLU	SGD

diverse in terms of image quality, contrast enhancement levels, and artifacts because it is only curated using keywords and metadata, and not by manual assessment.

From the UNN dataset, a subset of 10 randomly selected volumes from 10 different and randomly selected patients with liver tumors were manually labeled with segmentation masks of the liver and metastatic regions by a clinician (co-author K.R.) to be used in our study. In addition, two volumes from a patient that had been treated with liver surgery to remove a metastatic liver segment were included. One volume was before the surgery, and one after the surgery. The study of these pre- and post-operative images is conducted as a use-case of cross-examination CBIR.

7. Data robustness

To ensure real life clinical benefits, automatic algorithms should be able to handle data from different sources. The data used in this manuscript has been collected from multiple different sites. The Dechatlon dataset consists of data from seven clinical sites, and the UNN dataset provides an additional dataset for increased diversity. By training and evaluating on images from multiple clinical sites, we work towards creating algorithms that are robust to shifts in the data.

8. Experiments

We present the results of the experimental evaluation of our proposed framework. All models were trained with a batch size of 32 and for 250 epochs. Optimization was carried out using stochastic gradient descent with momentum=0.9, weight decay=0.0001, and learning rate=0.05 * batch size/256, as used in the SimSiam framework (Chen and He, 2021). As in previous works (Chen et al., 2020; Chen and He, 2021), a ResNet50 (He et al., 2016) was used as the feature extractor, with the output of the average pooling layer as the final representation. The ResNet50 is a convolutional neural network consisting of 48 convolutional layers, one max pooling layer, and one average pooling layer. The activation function used in the ResNet50 is the ReLU activation function (Glorot et al., 2011). The important advantage of the ResNet family of convolutional neural networks is the use of skip connections between computational layers. This improves the gradient flow in the neural networks which enables deeper networks with higher capability to be trained. The ResNet50 is one of the most widely used convolutional neural networks for computer vision tasks. For both the KNN classifier and the MAP we set K=5. Code is available at https://github.com/Wickstrom/clinical-self-supervised-CBIR-ct-liver.git. Table 1 gives an overview of several important components used in the proposed framework.

Table 2
Mean and std of mean average precision, knn accuracy and relevance rank score across 20 training runs on the test images from the Decathlon dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicate a statistically significant improvement at a significance level of 0.05. Significance was determined using a permutation test (Welch, 1990).

Pretraining	MAP	ACC	RR
IN	79.4	80.3	5.00
IN + SS (baseline)	87.5 ± 1.0	88.7 ± 0.8	**14.7 ± 3.1**
IN + SS (ours)	**88.7 ± 0.8**	**91.2 ± 1.1**	**14.8 ± 2.1**

Table 3
Mean and std of mean average precision, knn accuracy and relevance rank score across 20 training runs on test images from the UNN dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicate a statistically significant improvement at a significance level of 0.05. Significance was determined using a permutation test (Welch, 1990).

Pretraining	MAP	ACC	RR
IN	80.7	83.0	4.34
IN + SS (baseline)	**83.7 ± 1.7**	86.2 ± 2.4	**18.4 ± 3.0**
IN + SS (ours)	**84.3 ± 1.7**	**87.8 ± 2.3**	**18.0 ± 3.5**

8.1. Quantitative results

Tables 2 and 3 present the MAP, accuracy of a 5NN classifier, and the RR on the test data from the Decathlon and UNN datasets. The results show that the proposed framework outperforms the standard self-supervised approach across most scores. Furthermore, self-supervised learning greatly improves upon simply using the feature extractor trained on the Imagenet dataset. Also, the improvements are transferable across datasets, as the feature extractors trained on the Decathlon data also leads to improved performance in the UNN data.

Figs. 6 and 7 present the evolution of MAP, accuracy of a 5NN classifier, and the RR on the test data from the Decathlon and UNN datasets across training. The plots highlight how the scores improve as training progresses and stabilizes. However, an interesting observation is that the MAP and KNN accuracy achieves its highest value earlier in training on the UNN dataset, which suggests that the feature extractor is starting to learn features specific to the Decathlon dataset. In future works, this could be addressed by introducing more regularization into the self-supervised training procedure.

8.2. Explaining representations — qualitative results

The relevance rank scores in Tables 2 and 3 show that the proposed framework utilizes liver features in the image to a larger degree than the baseline approaches. However, the scores are far from perfect, which means that other parts of the image are also being used. Also, the feature extractor that is only trained on the Imagenet dataset has a very low relevance rank score, meaning that it is putting little attention on the liver. All of these observations can be investigated through XAI. In this section, we illuminate these observations through a new explainability analysis for CBIR by leveraging the RELAX framework that was described in Section 4.1. We show 4 qualitative examples, where the first example shows explanations for the feature extractor trained using Imagenet, and the remaining three examples shows explanations

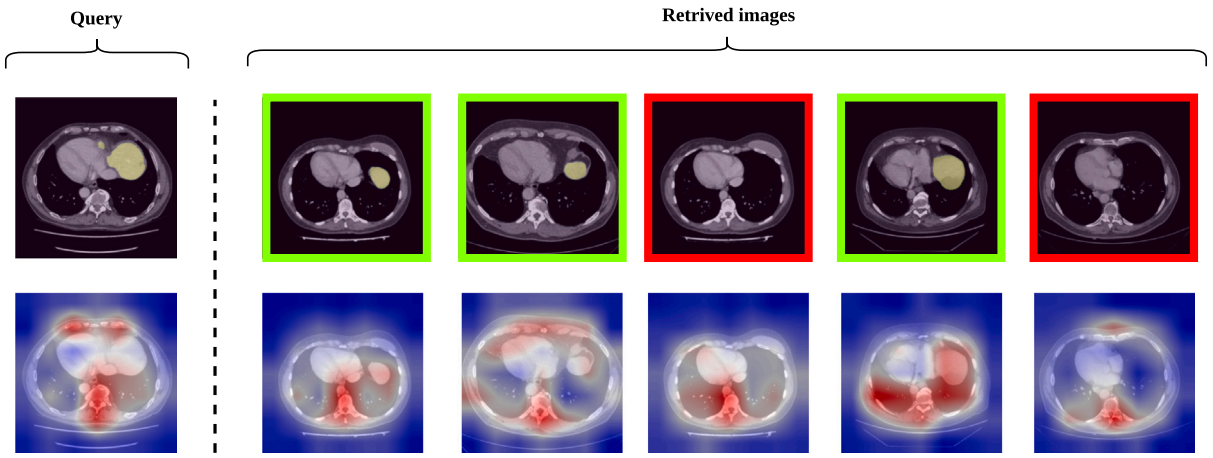

Fig. 8. Example (1): CBIR example from Decathlon dataset with feature extractor pretrained on Imagenet dataset. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. Some of the retrieved images do not contain the liver, and the explainability analysis shows that the feature extractor is focusing on the spine and rib cage instead of the organs. This information is important to understand why non-relevant images are retrieved, and would not be available without the explainability analysis.

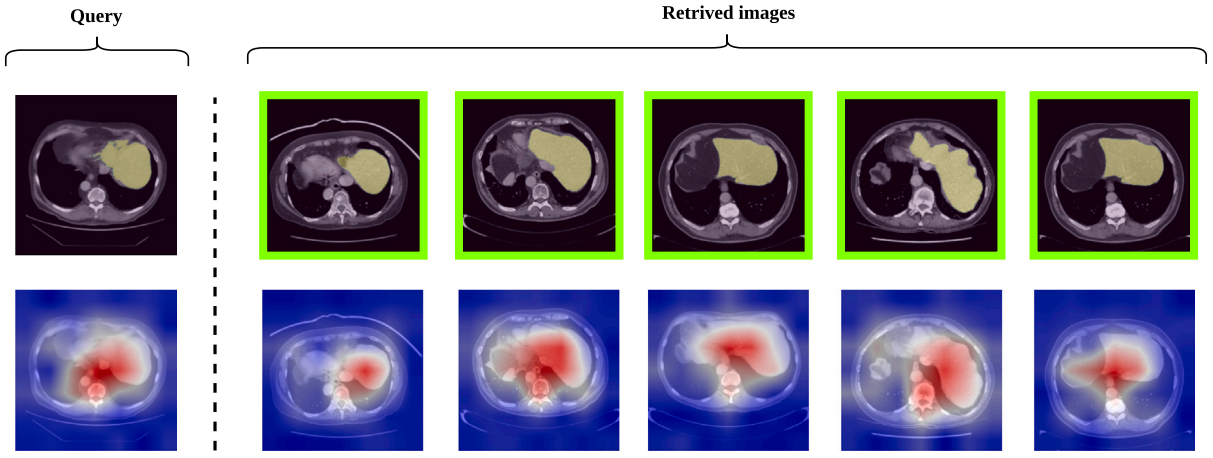

Fig. 9. Example (2): CBIR example from Decathlon dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver, and the explainability analysis shows that the feature extractor is focusing on the liver.

for the feature extractor trained using the proposed framework. In all examples, we show a query from the test set and the 5 retrieved images by the CBIR system. Additionally, we show the explanation for the query and retrieved images. The explanation show which features in the input are the most important for the representation of the image, where important pixels are highlighted in red and non-important pixels in blue.

Example 1: the feature extractor pretrained on Imagenet focuses on hard edges such as the spine. Fig. 8 displays an example where 2 of the 5 the retrieved images do not contain parts of the liver. When inspecting the explanations, it is clear that the feature extractor is not focusing on the liver, but rather on the tailbone. We hypothesize that since the feature extractor has never been presented with CT images, it utilizes prominent features with hard edges such as the spine, as opposed to organs with softer boundaries. The behavior discovered in this example is important, as it might also result in unexpected or poor retrievals for other queries.

Example 2: the feature extractor trained using the proposed framework focuses on liver features. Fig. 9 shows an example where all the retrieved images contain liver. Additionally, it is evident that the feature extractor is putting more emphasis on the liver for all the images, which illustrates how the proposed self-supervised framework has enabled the feature extractor to focus on clinically relevant features.

Example 3: the feature extractor trained using the proposed framework uses features from organs that often co-occur with the liver. Fig. 10 displays an example where the CBIR system retrieves 5 images that contain the liver, but where the explainability analysis shows that it not focusing on part of the images where the liver is present. Instead, it puts attention on the kidneys, which are quite prominent in all images. The kidneys often occur together with the liver in many CT images, and it also has similar pixel intensities as the liver (in terms of Hounsfield units). Therefore, it is not surprising that the feature extractor has learned to utilize both liver and kidney features, which also explains the behavior in this example. Such insights would not be obtainable without conducting the explainability analysis.

Example 4: the feature extractor trained using the proposed framework focuses on liver features, also for images from a different dataset. Lastly, Fig. 11 shows an example from the UNN dataset. This example illustrates that also on this new and unseen dataset, the feature extractor is basing the representation of these images features associated with the liver.

8.3. Case study: Cross-examination CBIR

A typical scenario in clinical practice is comparing a newly conducted examination with one ore more previous examinations. For

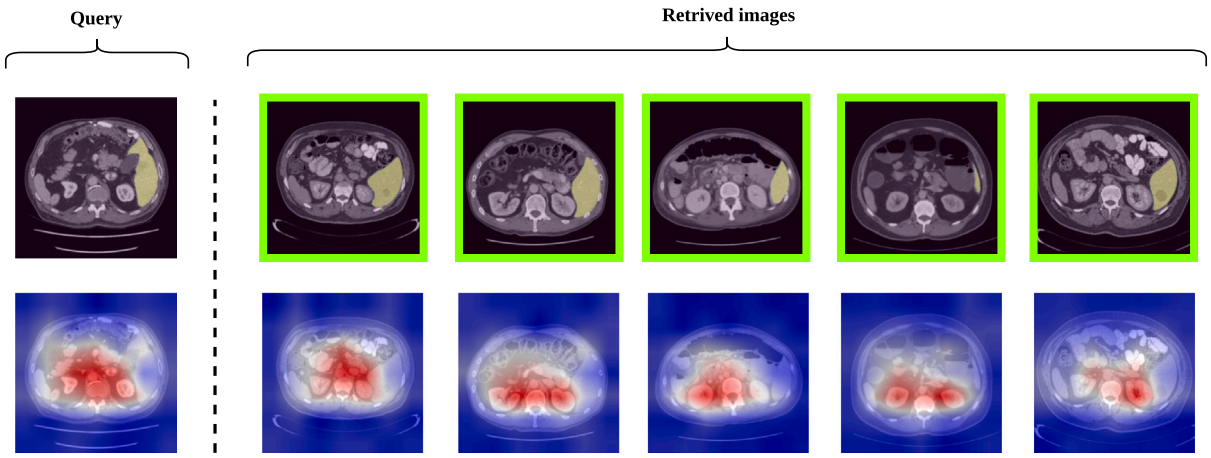

Fig. 10. Example (3): CBIR example from Decathlon dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver, but the explainability analysis reveals that the focus is on the kidneys, not the liver.

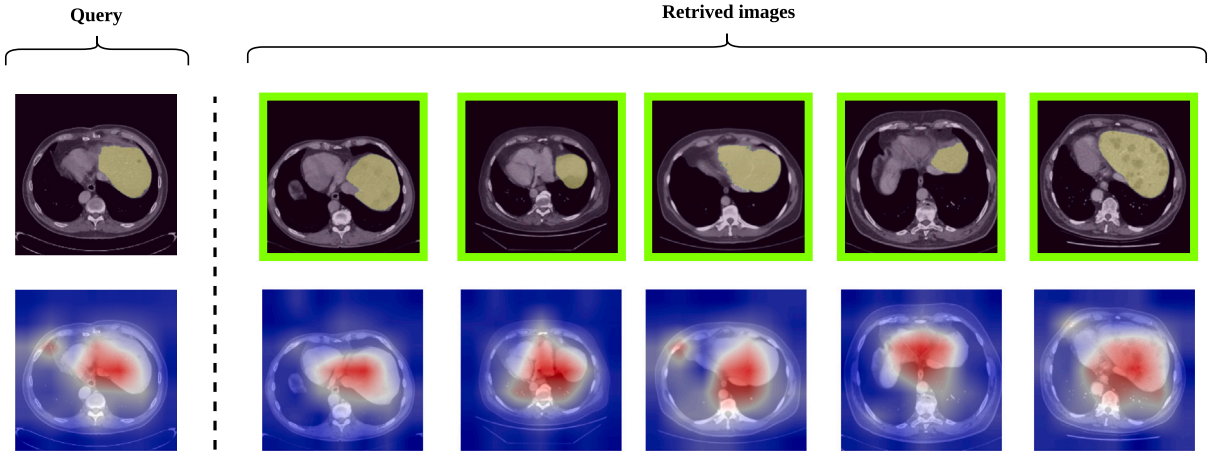

Fig. 11. Example (4): CBIR example from UNN dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver and the feature extractor is focusing on the liver, which illustrates that the feature extractor trained on the Decathlon dataset transfers well to the UNN dataset.

instance, one might want to compare a particular slice from the new examination with a selection of slices from one or several previous examinations. Such a comparison can help physicians understand how a disease has progressed since the previous examination, such as the development of liver metastasis. But when conducting such a comparison, the physician must manually inspect the new examination, and potentially several previous examinations. The CT scans are often taken with different settings across examinations, and it is therefore not possible to simply select the same slice from different examinations, as this can image completely different parts of the patient. A precise and reliable CBIR system could make such a cross-examination more efficient, by automating the retrieval process for the physician.

Fig. 12 displays an example of such a cross-examination. The query is selected from a recent examination, and the retrieved images are from the previous examination of the same patient. This patient is from the UNN dataset and was selected since liver metastasis has been developed between the two examinations. The query was selected by an experienced physician (co-author K.R.), which also selected five images to examine from the previous examination. Ideally, the CBIR system should align well with the image selected by the physician. In this example, the CBIR system produces a successful retrieval, as it identifies the same images as the physician. However, an interesting observation is that the images retrieved by the CBIR system are not sorted in the same manner as the physician's retrievals. This deviation might be due to the CBIR system being trained on single slices without a sense of spatial coherence. Future works could address this by incorporating neighboring samples as positive pairs in the self-supervised training.

8.4. Comparison with non-deep learning feature extractors

We compare a feature extractor trained using our proposed self-supervised approach with non-deep learning feature extraction methods. We consider Gabor filters (Granlund, 1978) as in prior works (Zhao et al., 2004; Lee et al., 2006) and histogram of oriented gradients (HOG) (Dalal and Triggs, 2005). For the Gabor filters we take inspiration from the prior work of Lee et al. (2006) and build a filter bank with 6 orientations, 4 scales, and 3×3 filter sizes. For each filtered image, we take the mean, variance, minimum value, and maximum value, which leads to a feature vector of size 96 for each image. For the HOG feature extraction we follow the prior work of Purojin Shamini (2018) and use 9×9 pixels per cell, and use 9 number of orientation bins. The results are shown in Tables 4 and 5 for the Decathlon and UNN data, respectively. For both datasets, the features produced by the deep learning architecture trained using our proposed self-supervised methodology clearly outperforms the previous approaches that use handcrafted features.

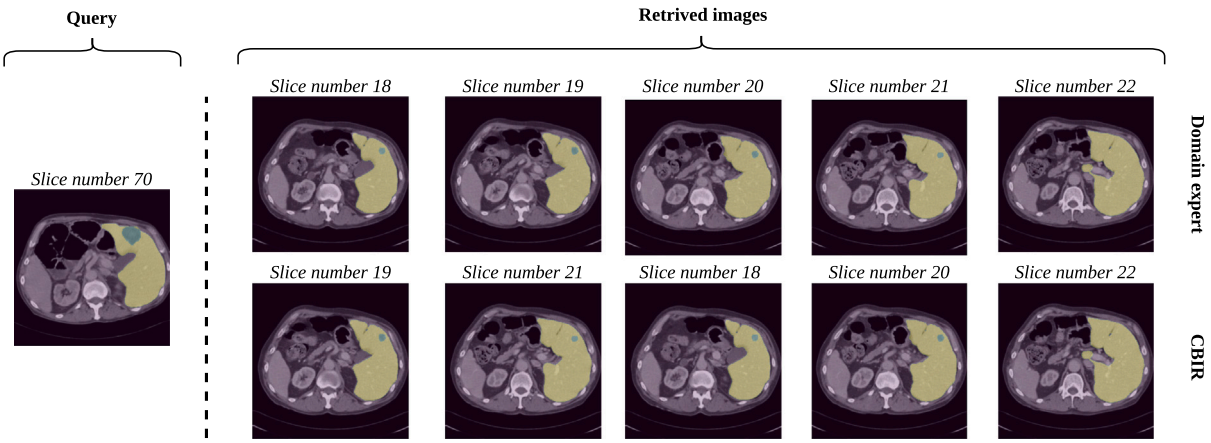

Fig. 12. An example of cross-examination CBIR. The query is from a recent examination, and the retrieved images are from a prior examination from the same patient. The goal of such a study is to investigate the development of liver metastasis. The query and retrieved images in the top row are selected by an experienced physician, and the bottom row are the retrieved images from the CBIR system. The CBIR successfully retrieves the same images as the physician, but lacks the spatial coherence to order the retrieved images.

Table 4
Mean and std of mean average precision and knn accuracy across 20 training runs on the test images from the Decathlon dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicate a statistically significant improvement at a significance level of 0.05. Significance was determined using a permutation test (Welch, 1990).

Feature extractor	MAP	ACC
Gabor	0.50	0.50
HOG	0.77	0.77
Ours	**89.1 ± 1.3**	**90.6 ± 1.4**

Table 5
Mean and std of mean average precision and knn accuracy across 20 training runs on the test images from the UNN dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicate a statistically significant improvement at a significance level of 0.05. Significance was determined using a permutation test (Welch, 1990).

Feature extractor	MAP	ACC
Gabor	0.49	0.49
HOG	0.78	0.75
Ours	**89.1 ± 1.3**	**90.6 ± 1.4**

9. Conclusion

We proposed a clinically motivated self-supervised framework for CBIR of CT liver images. Our proposed framework exploits the properties of the liver to learn more clinically relevant features, which leads to improved performance. Moreover, we leverage the RELAX framework to provide the first representation learning explainability analysis in the context of CBIR of CT liver images. Our analysis provides new insights into the feature extraction process and shows how self-supervised learning can provide feature extractors that extract more clinically relevant features compared to feature extractors trained on non-CT liver images. Our experimental evaluation also shows how the proposed framework generalizes to new datasets, and we present a clinically relevant user study. In future works, we intend to investigate how the proposed approach can be extended to extract features specific to other organs based on clipping strategies catered specifically to the desired organ. We believe that the proposed framework can play an essential role in constructing reliable CBIR that can effectively utilize unlabeled data.

CRediT authorship contribution statement

Kristoffer Knutsen Wickstrøm: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Eirik Agnalt Østmo:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Keyur Radiya:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Karl Øyvind Mikalsen:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Michael Christian Kampffmeyer:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Robert Jenssen:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

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

Parts of the data is publicly available while parts are not allowed to be shared.

Acknowledgments

This work was supported by The Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme [grant number 309439] and Consortium Partners; RCN FRIPRO [grant number 315029]; RCN IKTPLUSS [grant number 303514]; and the UiT Thematic Initiative.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2022. The medical segmentation decathlon. Nature Commun. 13 (1), http://dx.doi.org/10.1038/s41467-022-30695-9.

Arras, L., Osman, A., Samek, W., 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Inf. Fusion 81, 14–40. http://dx.doi.org/10.1016/j.inffus.2021.11.008.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M., 2021. Big self-supervised models advance medical image classification. In: International Conference on Computer Vision. pp. 3478–3488.

Ballerini, L., Li, X., Fisher, R.B., Rees, J., 2010. A query-by-example content-based image retrieval system of non-melanoma skin lesions. In: Medical Content-Based Retrieval for Clinical Decision Support. Springer Berlin Heidelberg, pp. 31–38. http://dx.doi.org/10.1007/978-3-642-11769-5_3.

Barata, C., Santiago, C., 2021. Improving the explainability of skin cancer diagnosis using CBIR. In: Medical Image Computing and Computer Assisted Intervention. Springer International Publishing, pp. 550–559.

Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.-W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.-A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.-H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.-K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.-H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B., 2023. The liver tumor segmentation benchmark (LiTS). Med. Image Anal. 84, 102680. http://dx.doi.org/10.1016/j.media.2022.102680.

Bravo, A.A., Sheth, S.G., Chopra, S., 2001. Liver biopsy. N. Engl. J. Med. 344 (7), 495–500. http://dx.doi.org/10.1056/nejm200102153440706.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. pp. 1877–1901.

Caron, M., Bojanowski, P., Joulin, A., Douze, M., 2018. Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision. pp. 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. In: Neural Information Processing Systems. pp. 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: International Conference on Computer Vision.

Chen, H., Gomez, C., Huang, C.-M., Unberath, M., 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. Npj Digit. Med. 5 (1), http://dx.doi.org/10.1038/s41746-022-00699-2.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Computer Vision and Pattern Recognition. pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607.

Chi, Y., Zhou, J., Venkatesh, S.K., Tian, Q., Liu, J., 2013. Content-based image retrieval of multiphase CT images for focal liver lesion characterization. Med. Phys. 40 (10), 103502.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 15 (141), 20170387. http://dx.doi.org/10.1098/rsif.2017.0387.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Computer Vision and Pattern Recognition. Vol. 1. pp. 886–893. http://dx.doi.org/10.1109/CVPR.2005.177.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186.

Dong, B., Collins, R., Hoogs, A., 2019. Explainability for content-based image retrieval. In: Computer Vision and Pattern Recognition Workshops. pp. 95–98.

Dong, N., Kampffmeyer, M., Voiculescu, I., 2021. Self-supervised multi-task representation learning for sequential medical images. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 779–794.

Doshi-Velez, F., Kim, B., 2017. Towards A rigorous science of interpretable machine learning. arXiv:1702.08608.

França, A.V.C., Valério, H.M.G., Trevisan, M., Escanhoela, C., Sevá-Pereira, T., Zucoloto, S., Martinelli, A., Soares, E.C., 2003. Fine needle aspiration biopsy for improving the diagnostic accuracy of cut needle biopsy of focal liver lesions. Acta Cytol. 47 (3), 332–336. http://dx.doi.org/10.1159/000326529.

Franceschi, J.-Y., Dieuleveut, A., Jaggi, M., 2019. Unsupervised scalable representation learning for multivariate time series. In: Neural Information Processing Systems. Vol. 32. pp. 4650–4661.

Gautam, S., Höhne, M.M.-C., Hansen, S., Robert Jenssen, M.K., 2022. Demonstrating the risk of imbalanced datasets in chest X-ray image-based diagnostics by prototypical relevance propagation. In: International Symposium on Biomedical Imaging.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Gordon, G., Dunson, D., Dudí k, M. (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol. 15, PMLR, Fort Lauderdale, FL, USA, pp. 315–323, URL: https://proceedings.mlr.press/v15/glorot11a.html.

Granlund, G.H., 1978. In search of a general picture processing operator. Comput. Graph. Image Process. 8 (2), 155–173. http://dx.doi.org/10.1016/0146-664x(78)90047-3.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent - A new approach to self-supervised learning. In: Neural Information Processing Systems. Vol. 33. pp. 21271–21284.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40 (2), 699–711.

Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M., 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Med. Image Anal. 78, 102385.

Haq, N.F., Moradi, M., Wang, Z.J., 2021. A deep community based approach for large scale content based X-ray image retrieval. Med. Image Anal. 68, 101847.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Computer Vision and Pattern Recognition. pp. 9726–9735.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 CVPR. pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.-C., 2023. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. J. Mach. Learn. Res. 24 (34), 1–11, URL: http://jmlr.org/papers/v24/22-0142.html.

Hu, B., Vasu, B., Hoogs, A., 2022. X-MIR: EXplainable medical image retrieval. In: Winter Conference on Applications of Computer Vision. WACV, pp. 1544–1554.

Jiang, M., Zhang, S., Metaxas, D.N., 2014. Detection of mammographic masses by content-based image retrieval. In: Machine Learning in Medical Imaging. pp. 33–41.

Jing, L., Vincent, P., LeCun, Y., Tian, Y., 2022. Understanding dimensional collapse in contrastive self-supervised learning. In: International Conference on Learning Representations.

Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S.T., Navab, N., 2021. Explaining COVID-19 and thoracic pathology model predictions by identifying informative input features. In: Medical Image Computing and Computer Assisted Intervention. pp. 391–401.

Kolesnikov, A., Zhai, X., Beyer, L., 2019. Revisiting self-supervised visual representation learning. In: IEEE Computer Vision and Pattern Recognition.

Kreft, B., 2001. Häufigkeit und bedeutung von kleinen fokalen leberläsionen in der MRT. In: Röfo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren. Vol. 173. No. 05. Georg Thieme Verlag KG, pp. 424–429. http://dx.doi.org/10.1055/s-2001-13340.

Lee, C.-C., Chen, S.-H., Tsai, H.-M., Chung, P.-C., Chiang, Y.-C., 2006. Discrimination of liver diseases from CT images based on gabor filters. In: 19th IEEE Symposium on Computer-Based Medical Systems. CBMS'06, pp. 203–206. http://dx.doi.org/10.1109/CBMS.2006.77.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A., 2018a. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imag. 2663–2674.

Li, Z., Zhang, X., Müller, H., Zhang, S., 2018b. Large-scale retrieval for medical image analytics: A comprehensive review. Med. Image Anal. 43, 66–84.

Marwaha, J.S., Kvedar, J.C., 2022. Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of AI. Npj Digit. Med. 5 (1), http://dx.doi.org/10.1038/s41746-022-00572-2.

Mirasadi, M.S., Foruzan, A.H., 2019. Content-based medical image retrieval of CT images of liver lesions using manifold learning. Int. J. Multimedia Inform. Retr. 8 (4), 233–240.

Monowar, M.M., Hamid, M.A., Ohi, A.Q., Alassafi, M.O., Mridha, M.F., 2022. AutoRet: A self-supervised spatial recurrent network for content-based image retrieval. Sensors.

Nam, D., Chapiro, J., Paradis, V., Seraphin, T.P., Kather, J.N., 2022. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. JHEP Rep. 4 (4), 100443. http://dx.doi.org/10.1016/j.jhepr.2022.100443.

Peng, T., Boxberg, M., Weichert, W., Navab, N., Marr, C., 2019. Multi-task learning of a deep K-nearest neighbour network for histopathological image classification and retrieval. In: Lecture Notes in Computer Science, pp. 676–684.

Petsiuk, V., Das, A., Saenko, K., 2018. RISE: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference. p. 151.

Plummer, B.A., Vasileva, M.I., Petsiuk, V., Saenko, K., Forsyth, D., 2020. Why do these match? Explaining the behavior of image similarity models. In: European Conference on Computer Vision. pp. 652–669.

Purojin Shamini, P.S.B., 2018. Automatic detection and classification technique for liver tumor in ct images. In: International Conference on Energy Efficient Technologies for Sustainability.

Quellec, G., Al Hajj, H., Lamard, M., Conze, P.-H., Massin, P., Cochener, B., 2021. ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. Med. Image Anal. 72, 102118.

Radiya, K., Joakimsen, H.L., Mikalsen, K.Ø., Aahlin, E.K., Lindsetmo, R.-O., Mortensen, K.E., 2023. Performance and clinical applicability of machine learning in liver computed tomography imaging: A systematic review. Eur. Radiol..

Ramalhinho, J., Tregidgo, H.F.J., Gurusamy, K., Hawkes, D.J., Davidson, B., Clarkson, M.J., 2021. Registration of untracked 2D laparoscopic ultrasound to CT images of the liver using multi-labelled content-based image retrieval. IEEE Trans. Med. Imaging 40 (3), 1042–1054.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R., 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126 (4), 552–564. http://dx.doi.org/10.1016/j.ophtha.2018.11.016.

Schulz, K., Sixt, L., Tombari, F., Landgraf, T., 2020. Restricting the flow: Information bottlenecks for attribution. In: International Conference on Learning Representations.

Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M., 2020. Interpretability-guided content-based medical image retrieval. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer-Verlag, Berlin, Heidelberg, pp. 305–314. http://dx.doi.org/10.1007/978-3-030-59710-8_30.

Singal, A.G., Mukherjee, A., Elmunzer, J.B., Higgins, P.D.R., Lok, A.S., Zhu, J., Marrero, J.A., Waljee, A.K., 2013. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am. J. Gastroenterol. 108 (11), 1723–1730. http://dx.doi.org/10.1038/ajg.2013.332.

Siradjuddin, I.A., Wardana, W.A., Sophan, M.K., 2019. Feature extraction using self-supervised convolutional autoencoder for content based image retrieval. In: International Conference on Informatics and Computational Sciences. pp. 1–5.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net. In: International Conference on Learning Representations Workshop.

Tapper, E.B., Lok, A.S.-F., 2017. Use of liver imaging and biopsy in clinical practice. In: Longo, D.L. (Ed.), N. Engl. J. Med. 377 (8), 756–768. http://dx.doi.org/10.1056/nejmra1610570.

Thomaides-Brears, H.B., Alkhouri, N., Allende, D., Harisinghani, M., Noureddin, M., Reau, N.S., French, M., Pantoja, C., Mouchti, S., Cryer, D.R.H., 2021. Incidence of complications from percutaneous biopsy in chronic liver disease: A systematic review and meta-analysis. Dig. Dis. Sci. 67 (7), 3366–3394. http://dx.doi.org/10.1007/s10620-021-07089-w.

Tian, Y., Chen, X., Ganguli, S., 2021. Understanding self-supervised learning dynamics without contrastive pairs. In: International Conference on Machine Learning. Vol. 139. pp. 10268–10278.

Tisch, C., Brencicova, E., Schwendener, N., Lombardo, P., Jackowski, C., Zech, W.-D., 2019. Hounsfield unit values of liver pathologies in unenhanced post-mortem computed tomography. Int. J. Legal Med. 1861–1867.

Tiyarattanachai, T., Apiparakoon, T., Marukatat, S., Sukcharoen, S., Yimsawad, S., Chaichuen, O., Bhumiwat, S., Tanpowpong, N., Pinjaroen, N., Rerknimitr, R., Chaiteerakij, R., 2022. The feasibility to use artificial intelligence to aid detecting focal liver lesions in real-time ultrasound: a preliminary study based on videos. Sci. Rep. 12 (1), http://dx.doi.org/10.1038/s41598-022-11506-z.

Truong, T., Mohammadi, S., Lenga, M., 2021. How transferable are self-supervised features in medical image classification tasks? In: Proceedings of Machine Learning for Health. Vol. 158. pp. 54–74.

Vasilakakis, M., Sovatzidi, G., Iakovidis, D.K., 2021. Explainable classification of weakly annotated wireless capsule endoscopy images based on a fuzzy bag-of-colour features model and brain storm optimization. In: Medical Image Computing and Computer Assisted Intervention. pp. 488–498.

Wang, K., Lu, X., Zhou, H., Gao, Y., Zheng, J., Tong, M., Wu, C., Liu, C., Huang, L., Jiang, T., Meng, F., Lu, Y., Ai, H., Xie, X.-Y., ping Yin, L., Liang, P., Tian, J., Zheng, R., 2018. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 68 (4), 729–741. http://dx.doi.org/10.1136/gutjnl-2018-316204.

Welch, W.J., 1990. Construction of permutation tests. J. Amer. Statist. Assoc. 85 (411), 693–698. http://dx.doi.org/10.1080/01621459.1990.10474929.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. 60, 101619.

Wickstrøm, K., Kampffmeyer, M., Mikalsen, K.Ø., Jenssen, R., 2022. Mixing up contrastive learning: Self-supervised representation learning for time series. Pattern Recognit. Lett. 155, 54–61. http://dx.doi.org/10.1016/j.patrec.2022.02.007.

Wickstrøm, K., Mikalsen, K., Kampffmeyer, M., Revhaug, A., Jenssen, R., 2021. Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series. IEEE J. Biomed. Health Inf. 25 (7), 2435–2444. http://dx.doi.org/10.1109/JBHI.2020.3042637.

Wickstrøm, K.K., Trosten, D.J., Løkse, S., Boubekki, A., Mikalsen, K.Ø., Kampffmeyer, M.C., Jenssen, R., 2023. RELAX: Representation learning explainability. Int. J. Comput. Vis. http://dx.doi.org/10.1007/s11263-023-01773-2.

Yasaka, K., Akai, H., Kunimatsu, A., Abe, O., Kiryu, S., 2018. Liver fibrosis: Deep convolutional neural network for staging by using gadoxetic acid–enhanced hepatobiliary phase MR images. Radiology 287 (1), 146–155. http://dx.doi.org/10.1148/radiol.2017171928.

Yoshinobu, Y., Iwamoto, Y., Han, X., Lin, L., Hu, H., Zhang, Q., Chen, Y.-W., 2020. Deep learning method for content-based retrieval of focal liver lesions using multiphase contrast-enhanced computer tomography images. In: International Conference on Consumer Electronics. pp. 1–4.

Zhao, C., Cheng, H., Huo, Y., Zhuang, T., 2004. Liver CT-image retrieval based on Gabor texture. In: International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 1. pp. 1491–1494.

Zheng, Y., Jiang, B., Shi, J., Zhang, H., Xie, F., 2019. Encoding histopathological WSIs using GNN for scalable Diagnostically Relevant Regions retrieval. In: Lecture Notes in Computer Science. pp. 550–558.

Zheng, G., Shi, L., Liu, J., Zhao, Y., Du, F., He, Y., Yang, X., Song, N., Wen, J., Gao, H., 2022. Registered trials of artificial intelligence conducted on chronic liver disease: A cross-sectional study on ClinicalTrials.gov. In: Yang, X. (Ed.), Dis. Markers 2022, 1–8. http://dx.doi.org/10.1155/2022/6847073.

Zhou, L.-Q., Wang, J.-Y., Yu, S.-Y., Wu, G.-G., Wei, Q., Deng, Y.-B., Wu, X.-L., Cui, X.-W., Dietrich, C.F., 2019. Artificial intelligence in medical imaging of the liver. World J. Gastroenterol. 25 (6), 672–682. http://dx.doi.org/10.3748/wjg.v25.i6.672.