Medical Image Analysis 91 (2024) 103042

Contents lists available at ScienceDirect

Medical Image Analysis

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

Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis

Ričards Marcinkevičs [a,1,*], Patricia Reis Wolfertstetter [b,c,1], Ugne Klimiene [a,1], Kieran Chin-Cheong [a], Alyssia Paschke [c], Julia Zerres [c], Markus Denzinger [b,c], David Niederberger [a], Sven Wellmann [c,d], Ece Ozkan [e,2], Christian Knorr [b,2], Julia E. Vogt [a,2]

[a] Department of Computer Science, ETH Zurich, Universitätstrasse 6, Zürich, 8092, Switzerland
[b] Department of Pediatric Surgery and Pediatric Orthopedics, Hospital St. Hedwig of the Order of St. John of God, University Children's Hospital Regensburg (KUNO), Steinmetzstrasse 1-3, Regensburg, 93049, Germany
[c] Faculty of Medicine, University of Regensburg, Franz-Josef-Strauss-Allee 11, Regensburg, 93053, Germany
[d] Division of Neonatology, Hospital St. Hedwig of the Order of St. John of God, University Children's Hospital Regensburg (KUNO), Steinmetzstrasse 1-3, Regensburg, 93049, Germany
[e] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, 02139, USA

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.7711412, https://github.com/i6092467/semi-supervised-multiview-cbm

MSC:
41A05
41A10
65D05
65D17

Keywords:
Interpretable machine learning
Multiview learning
Concepts
Classification
Pediatric appendicitis
Ultrasound imaging

ABSTRACT

Appendicitis is among the most frequent reasons for pediatric abdominal surgeries. Previous decision support systems for appendicitis have focused on clinical, laboratory, scoring, and computed tomography data and have ignored abdominal ultrasound, despite its noninvasive nature and widespread availability. In this work, we present interpretable machine learning models for predicting the diagnosis, management and severity of suspected appendicitis using ultrasound images. Our approach utilizes concept bottleneck models (CBM) that facilitate interpretation and interaction with high-level concepts understandable to clinicians. Furthermore, we extend CBMs to prediction problems with multiple views and incomplete concept sets. Our models were trained on a dataset comprising 579 pediatric patients with 1709 ultrasound images accompanied by clinical and laboratory data. Results show that our proposed method enables clinicians to utilize a human-understandable and intervenable predictive model without compromising performance or requiring time-consuming image annotation when deployed. For predicting the diagnosis, the extended multiview CBM attained an AUROC of 0.80 and an AUPR of 0.92, performing comparably to similar black-box neural networks trained and tested on the same dataset.

1. Introduction

Appendicitis is one of the most frequent causes of abdominal pain resulting in hospital admissions of patients under 18 (Wier et al., 2013). The diagnosis can be challenging and relies on a combination of clinical, laboratory and imaging parameters (Saverio et al., 2016). Despite extensive research, no specific and practically useful biomarkers for the early detection of appendicitis have been identified (Acharya et al., 2016; Kiss et al., 2021). Epidemiologically and clinically, there are two forms of appendicitis: uncomplicated (subacute/exudative, phlegmonous) and complicated (gangrenous, perforated) (Andersson, 2006; Bhangu et al., 2015; Kiss et al., 2021). Management forms include surgery as the standard method (Saverio et al., 2016; Gorter et al., 2016) or conservative therapy (Andersson, 2006; Svensson et al., 2012, 2015; Gorter et al., 2016; CODA Collaborative, 2020).

Typical imaging modalities for suspected pediatric appendicitis include ultrasonography (US), magnetic resonance imaging (MRI), and computed tomography (CT). US has become the primary choice due to widespread availability, lack of radiation, and improvements in resolution over the past years (Park et al., 2011). Repeated US examinations, including B(rightness)-mode and Doppler, during the observation phase

* Corresponding author.
E-mail addresses: ricardsm@inf.ethz.ch (R. Marcinkevičs), patricia.reiswolfertstetter@barmherzige-regensburg.de (P. Reis Wolfertstetter), julia.vogt@inf.ethz.ch (J.E. Vogt).
[1] These authors have contributed equally and share first authorship.
[2] These authors have contributed equally and share last authorship.

https://doi.org/10.1016/j.media.2023.103042
Received 30 March 2023; Received in revised form 10 November 2023; Accepted 20 November 2023
Available online 23 November 2023

can improve diagnostic accuracy and help identify disease progression (Dingemann and Ure, 2012; Ohba et al., 2016; Gorter et al., 2016).

Extensive research has been conducted on utilizing machine learning (ML) models to diagnose and manage patients with suspected appendicitis (Hsieh et al., 2011; Deleger et al., 2013; Reismann et al., 2019; Aydin et al., 2020; Akmese et al., 2020; Stiel et al., 2020; Rajpurkar et al., 2020; Marcinkevics et al., 2021; Roig Aparicio et al., 2021; Xia et al., 2022). In brief, most models either utilize simple clinical and laboratory data (Hsieh et al., 2011; Aydin et al., 2020; Akmese et al., 2020; Xia et al., 2022), rely on hand-crafted US annotations (Reismann et al., 2019; Stiel et al., 2020; Marcinkevics et al., 2021; Roig Aparicio et al., 2021), or require more expensive and invasive imaging modalities, such as CT (Rajpurkar et al., 2020). Despite having lower sensitivity and specificity than CT, US has been advocated as the preferred imaging modality for diagnosing acute appendicitis due to the absence of ionizing radiation and cost-effectiveness (Mostbeck et al., 2016). Although promising and practical, fully automated analysis of abdominal US images in this context remains an under-explored approach.

US imaging gives natural rise to multiview and multimodal data (Wang et al., 2020; Qian et al., 2021). For instance, the risk of breast cancer may be assessed based on multiview and multimodal US images of lesions. More generally, multiview learning (Xu et al., 2013) concerns itself with the data comprising multiple views, essentially feature subsets, of the same source object. Additionally, multimodal learning (Baltrušaitis et al., 2019) studies models combining, or fusing, multiple heterogeneous modalities, e.g. images and text. Both research directions have experienced renewed interest in the light of contrastive and self-supervised learning (Tian et al., 2020; von Kügelgen et al., 2021) and generative modeling (Suzuki and Matsuo, 2022).

Interpretable machine learning has emerged as an active research direction, (Doshi-Velez and Kim, 2017; Rudin, 2019), with interpretability argued to be an essential model design principle for high-stakes application domains, such as healthcare. One recently re-explored approach is prediction based on high-level and human-understandable concepts (Kumar et al., 2009; Lampert et al., 2009; Koh et al., 2020) or attributes. Most frameworks for concept-based prediction require auxiliary supervision in the form of high-level semantic features during training. Typically, two models are trained, as, for instance, in concept bottleneck models (CBM) (Koh et al., 2020): (i) one mapping from the explanatory variables to the given concepts and (ii) another predicting the target variable based on the previously predicted concept values. Such concept-based models are deemed interpretable since concepts can be inspected alongside the final model outputs and perceived as "explanations". Additionally, as opposed to classical multitask learning, a human user can intervene and interact with the model at test time by editing concept predictions and affecting downstream output. Beyond the restricted supervised setting mentioned earlier, there have been several efforts to learn semantically meaningful and identifiable representations when the concepts are not given explicitly (Khemakhem et al., 2020; Taeb et al., 2022).

This work presents the first effort at leveraging ML to predict diagnosis, management, and severity in pediatric patients with suspected appendicitis *directly* from abdominal US images, an imaging modality frequently used in daily clinical practice. To this end, our models utilize interpretable concept-based classification approach due to its potential acceptance among clinicians and investigate the trade-off between interpretability and predictive performance. Furthermore, we propose extensions of the concept bottleneck models (Koh et al., 2020) to improve their scalability to real-world medical imaging data, contributing to the recent works identifying and addressing the limitations of concept-based models (Mahinpei et al., 2021; Margeloiu et al., 2021; Sawada and Nakamura, 2022; Marconato et al., 2022). Specifically, we extend conventional CBMs (i) to the multiview classification setting and (ii) propose a semi-supervised representation learning approach to overcome the limitations of incomplete concept sets, i.e. when the given set of concepts does not capture the entire predictive relationship between the images and labels, making it challenging to achieve high predictive performance. The presented generalization of the CBMs to multiple views and incomplete concept sets is summarized in Fig. 1. It is not restricted to the considered use case of pediatric appendicitis and ultrasound and can be applied to other multiview and multimodal medical imaging datasets.

2. Materials and methods

2.1. Dataset

In our retrospective analysis, we examined data from a cohort of 579 children and adolescents (aged 0–18 years) admitted as inpatients to the Department of Pediatric Surgery and Pediatric Orthopedics at the tertiary Children's Hospital St. Hedwig in Regensburg, Germany between January 1, 2016, and December 31, 2021, with suspected appendicitis. Our study builds and expands upon the previous analysis of a smaller cohort of patients, published by Marcinkevics et al. (2021).

We utilized the hospital's database to collect retrospective data, including (potentially) multiple abdominal B-mode ultrasound images for each patient (totaling 1709 images). The number of views per subject ranges from 1 to 15; the images depict various regions of interest, such as the abdomen's right lower quadrant (RLQ), appendix, intestines, lymph nodes, and reproductive organs (Fig. 2). Ultrasound images from admission and, if available, initial clinical course were retrieved using the software Clinic WinData/E&L. For surgical patients, US images from the preoperative clinical course were also included. The images were acquired on Toshiba Xario and Aplio XG machines using Toshiba 6 MHz Convex and 12 MHz Linear transducers. For each subject, all images relevant to the findings from Table 2 were included. Images of the organs unrelated to appendicitis, such as the liver or spleen, were excluded from the dataset. We also retrieved information encompassing laboratory tests, physical examination results, clinical scores, such as Alvarado (AS) and pediatric appendicitis (PAS) scores (Alvarado, 1986; Samuel, 2002; RSGobotWMR, 2020). AS and PAS were utilized due to the widespread use by pediatricians and pediatric surgeons for the risk stratification of children and adolescents with abdominal pain (Dingemann and Ure, 2012). Last but not least, we collected expert-produced ultrasonographic findings represented by categorically-valued features. A subset of the latter was identified as high-level concepts relevant to decision support (Table 2). For patients treated operatively, surgical and histological parameters were recorded.

The subjects were labeled w.r.t. three target variables: (i) diagnosis (*appendicitis* vs. *no appendicitis*), (ii) management (*surgical* vs. *conservative*), and (iii) severity (*complicated* vs. *uncomplicated or no appendicitis*). The diagnosis was confirmed histologically in the patients who underwent appendectomy. Subjects treated conservatively were labeled as having appendicitis if their appendix diameter was at least 6mm and either AS or PAS were at least 4. Note that the labeling criterion above is only a proxy for the ground-truth disease status. AS and PAS help exclude children with no appendicitis (RSGobotWMR, 2020), whereas the addition of the US information on the enlarged appendix has been shown to increase the positive predictive value (Gendel et al., 2011; Dingemann and Ure, 2012). This labeling criterion has already been utilized in the previous analyses of the data from an overlapping patient cohort (Marcinkevics et al., 2021; Roig Aparicio et al., 2021). Marcinkevics et al. (2021) present a more detailed exploration to justify it. The management label reflects the decision made by a senior pediatric surgeon based on clinical, laboratory and US data. For the severity, complicated appendicitis includes cases with abscess formation, gangrene, or perforation.

Note that the analysis below utilizes only ultrasound images and findings extracted from them. Our goal was to explore US image analysis and its benefits for predictive models for pediatric appendicitis.

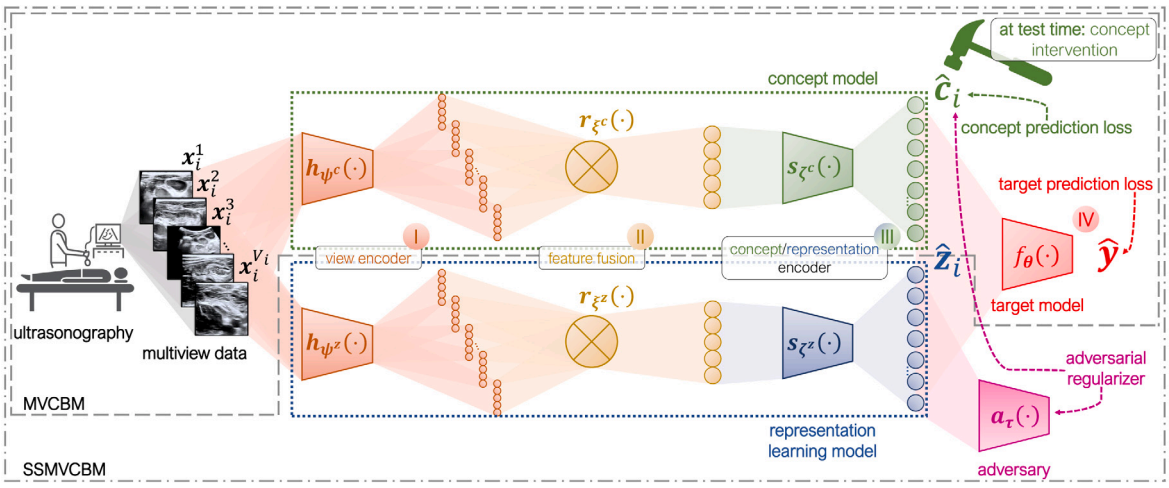

Fig. 1. Schematic summary of the proposed multiview concept bottleneck model (MVCBM) and its semi-supervised extension (SSMVCBM). (I) Multiview ultrasound images are mapped to features using a shared encoder neural network; (II) features are aggregated across the views; (III) high-level human-understandable concepts and representations are predicted based on the aggregated features; (IV) using concepts and representations, the target prediction is made. The MVCBM only includes view encoding, fusion, and concept prediction, whereas the SSMVCBM also performs representation learning. During training, in addition to the target prediction loss, the MVCBM is supervised by the concept prediction loss. The SSMVCBM is further penalized by an adversarial regularizer encouraging statistical independence between predicted concepts and representations.

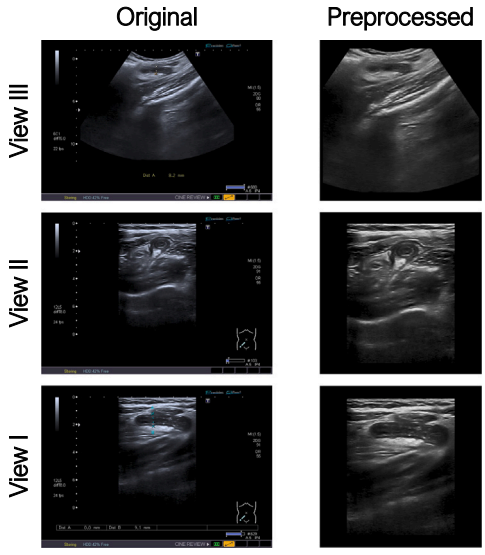

Fig. 2. An example of multiple US images acquired from a single patient from the pediatric appendicitis dataset. For this patient, views I and II correspond to longitudinal and transverse sections of the appendix, respectively; view III depicts the reaction in tissue surrounding the appendix. Original images (*left*) contain graphical interface elements and expert-made markers, whereas preprocessed images (*right*) have been inpainted, cropped, and padded.

Nevertheless, we publicize the entire dataset, including modalities other than imaging. Table 1 and Appendix A.1 provide an overview of the dataset used in the final analysis. Appendix A contains a more comprehensive description of the dataset and its acquisition.

2.1.1. Data preprocessing

Prior to model development and evaluation, pre-processing was performed on B-mode ultrasound images to eliminate undesired variability. The study being retrospective, ultrasonograms were collected as per clinical routine, and therefore, original images contained graphical user interface elements, markers, distance measurements, and other annotations. We employed a generative inpainting model DeepFill (Yu et al., 2018), to mask and fill such objects. Subsequently, images were resized to 400 × 400 px^2 dimensions using zero padding when needed.

Table 1
The contingency table of the pediatric appendicitis dataset of the management by severity stratified by the diagnosis.

	Diagnosis: *appendicitis*		
Management	Severity		
	complicated	*uncomplicated*	Total
surgical	97	135	232
conservative	0	151	151
Total	97	286	383
	Diagnosis: *no appendicitis*		
Management	Severity		
	complicated	*uncomplicated*	**Total**
surgical	0	2	2
conservative	0	194	194
Total	0	196	196

Finally, contrast-limited histogram equalization (CLAHE) was applied, and pixel intensities were normalized to the range of 0 and 1. Fig. 2 shows an example of the multiple US views acquired from a single subject from our cohort before and after preprocessing.

2.2. Problem setting and notation

Throughout the remaining sections, we will assume the following setting and notation. Consider a dataset comprising N triples $\left(\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}, \boldsymbol{c}_i, y_i\right)$, for $1 \leq i \leq N$, with view sequences $\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}$, concept vectors $\boldsymbol{c}_i \in \mathbb{R}^K$ provided at training time, and labels y_i. Note that the number of views $V_i \geq 1$ may vary across data points $1 \leq i \leq N$. We will concentrate on the scenario where all views can be preprocessed and rescaled into the same dimensionality. Nevertheless, our approach can be extended to heterogeneous data types.

Motivated by medical imaging applications, we focus on the data exhibiting characteristics described informally below. (i) *Partial observability*: not all concepts are identifiable from all views. (ii) *View homogeneity*: most views contain a considerable amount of shared information and are visually similar. (iii) *View ordering*: views belonging to the same data point may be loosely ordered, e.g. spatially, temporally, or based on their importance for predicting the label. These properties are inspired by the multiview ultrasound dataset explored in our experiments and support some design choices described below.

2.3. Multiview concept bottleneck models

Below, we present a novel approach that extends the concept bottleneck models (Koh et al., 2020) to the multiview classification scenario. We refer to this extension as the multiview concept bottleneck model (MVCBM) hereon. A schematic overview of the MVCBM architecture is shown in Fig. 1, while the model's forward pass is specified by Eqs. (1a)–(1d). In brief, MVCBM comprises four modules: (i) per-view feature extraction; (ii) feature fusion; (iii) concept prediction, and (iv) label prediction.

To address scenarios where the set of concepts provided is incomplete, aka insufficient, either due to the lack of domain knowledge or the cost of acquiring additional annotation, we have also developed a semi-supervised variant of the MVCBM, referred to as semi-supervised MVCBM (SSMVCBM). This approach not only utilizes the given concepts but also learns an independent representation predictive of the label. Note that this extension will be described in the later sections.

For data point $1 \leq i \leq N$, a forward pass of the multiview concept bottleneck is given by the following equations:

(i) Feature extraction:

$$\boldsymbol{h}_i^v = \boldsymbol{h}_{\boldsymbol{\psi}}\left(\boldsymbol{x}_i^v\right), \; 1 \leq v \leq V_i, \tag{1a}$$

(ii) Feature fusion:

$$\bar{\boldsymbol{h}}_i = \boldsymbol{r}_{\boldsymbol{\xi}}\left(\left\{\boldsymbol{h}_i^v\right\}_{v=1}^{V_i}\right), \tag{1b}$$

(iii) Concept prediction:

$$\hat{\boldsymbol{c}}_i = \boldsymbol{s}_{\boldsymbol{\zeta}}\left(\bar{\boldsymbol{h}}_i\right), \tag{1c}$$

(iv) Label prediction:

$$\hat{y}_i = f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{c}}_i\right), \tag{1d}$$

where Latin letters correspond to functions and variables and Greek letters denote learnable parameters. Observe that parameters $\boldsymbol{\phi} = \{\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\zeta}\}$ define the concept model $\boldsymbol{g}_{\boldsymbol{\phi}}(\cdot)$ mapping a multiview feature sequence to the predicted concept values; whereas $f_{\boldsymbol{\theta}}(\cdot)$ is the target model, linking concepts and labels. Thus, similar to the vanilla concept bottleneck, MVCBM's forward pass can be rewritten as $\hat{y}_i = f_{\boldsymbol{\theta}}\left(\boldsymbol{g}_{\boldsymbol{\phi}}\left(\left\{\boldsymbol{h}_i^v\right\}_{v=1}^{V_i}\right)\right)$. In the following paragraphs, we detail each of the steps in Eq. (1).

Feature extraction. Given an ordered view sequence $\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V_i}$, we first encode each view into a lower-dimensional representation, as in Eq. (1a). We employ a *shared* encoder neural network, denoted by $\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$. Weight sharing is justified by the view homogeneity and could be helpful in smaller datasets with high missingness of views. On the other hand, in multimodal datasets, the dissimilarities between images acquired from the same subject are significant and consistent. In this scenario, it may be preferable to train a dedicated encoder for each modality to learn modality-specific features. In practice, it may be prudent to use a pretrained model to initialize $\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$, e.g. the use of ResNet and VGG architectures pretrained on natural images is standard for medical imaging applications (Cheplygina, 2019). As a result, we obtain a sequence of view-specific features.

Feature fusion. To accommodate multiple views, we need to fuse, i.e. aggregate, the view-specific features within the model, as in Eq. (1b). MVCBM follows a *hybrid* fusion approach (Baltrušaitis et al., 2019): rather than concatenating views at the input level (*early fusion*) or training an ensemble of view-specific models (*late fusion*); we aggregate intermediate view-specific features $\boldsymbol{h}_i^v$ from the previous step within a single neural network. Although there are many viable fusion functions, in our context, the fusion must handle varying numbers of views per data point. As a naive approach, we consider arithmetic mean across the views $\bar{\boldsymbol{h}}_i = \frac{1}{V_i}\sum_{v=1}^{V_i} \boldsymbol{h}_i^v$ (Havaei et al., 2016).

More generally, in Eq. (1b) $\bar{\boldsymbol{h}}_i$ denotes the fused feature vector and $\boldsymbol{r}_{\boldsymbol{\xi}}(\cdot)$ is the fusion function with parameters $\boldsymbol{\xi}$. Considering partial observability of the concepts and ordering of the views, we, in addition, investigate aggregation via a *learnable* function. Similar to Ma et al. (2019), who utilize this trick in multiview 3D shape recognition, we combine view-specific representations via a long short-term memory (LSTM) network. In particular, we set the aggregated representation $\bar{\boldsymbol{h}}_i$ to the last hidden state of the view sequence, i.e. at step V_i. Note that both averaging and LSTM can handle varying numbers of views. Nevertheless, there are other options for $\boldsymbol{r}_{\boldsymbol{\xi}}(\cdot)$, e.g. Hadamard product or weighted average, the investigation of which we leave for future work.

Concept and label prediction. The last two steps in Eqs. (1c)–(1d) are similar to the vanilla concept bottleneck. First, we predict concepts $\hat{\boldsymbol{c}}_i$ based on the fused representation $\bar{\boldsymbol{h}}_i$, using a concept encoder network $\boldsymbol{s}_{\boldsymbol{\zeta}}(\cdot)$ parameterized by $\boldsymbol{\zeta}$. Note that the choice of activation functions at the output of $\boldsymbol{s}_{\boldsymbol{\zeta}}(\cdot)$ depends on the type of concepts and should be adapted to whether an individual concept is categorically or continuously valued. The vector $\hat{\boldsymbol{c}}_i$ is then used as an input to the target model $f_{\boldsymbol{\theta}}(\cdot)$, predicting the label $\hat{y}$. The output activation should be chosen based on the downstream task, which can be, for example, classification or regression.

Loss function and optimization. The parameters of vanilla CBMs can be optimized using independent, sequential and joint procedures (Koh et al., 2020). In this work, we focus on the sequential and joint approaches since they offer a more balanced trade-off between predictive performance and intervenability, as shown experimentally by Koh et al. (2020).

In the *sequential* training, we first optimize the concept model parameters:

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \sum_{i=1}^{N}\sum_{k=1}^{K} w_i^t w_i^{c_k} \mathcal{L}^{c_k}(\hat{c}_{i,k}, c_{i,k}), \tag{2}$$

where $\mathcal{L}^{c_k}(\cdot,\cdot)$ is the loss function for the kth concept, e.g. one could use the cross-entropy for categorically valued and squared error for a continuously valued concept, and $c_{i,k}$ refers to the value of the kth concept for the ith data point.

Additionally, to address potential imbalances in the concept distributions and sparsity of specific concept-target combinations, we have introduced weights $w_i^{c_k}$ for the kth concept and w_i^t for the target variable of the ith point, s.t. $\sum_{i=1}^N \sum_{k=1}^K w_i^{c_k} = 1$ and $\sum_{i=1}^N w_i^t = 1$. In practice, these weights can be set to the normalized inverse counts of samples in the corresponding variable classes, i.e. $w_i^t \propto 1/\sum_{j=1}^N \mathbf{1}_{\{y_j = y_i\}}$ and $w_i^{c_k} \propto 1/\sum_{j=1}^N \mathbf{1}_{\{c_{j,k} = c_{i,k}\}}$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. However, other sample weighting schemes are viable.

Next, parameters $\hat{\boldsymbol{\phi}}$ are frozen, and the parameters of the target model $f_{\boldsymbol{\theta}}$ are optimized:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} w_i^t \mathcal{L}^t\left(f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{c}}_i\right), y_i\right), \tag{3}$$

where $\mathcal{L}^t(\cdot,\cdot)$ is the loss function for the target task, and $\hat{\boldsymbol{c}}_i$ are predictions made by the frozen concept model $\boldsymbol{g}_{\hat{\boldsymbol{\phi}}}(\cdot)$.

For the *joint* training, we combine the loss functions from Eqs. (2) and (3) into a single objective:

$$\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} w_i^t \mathcal{L}^t(\hat{y}_i, y_i) + \alpha \sum_{i=1}^{N}\sum_{k=1}^{K} w_i^t w_i^{c_k} \mathcal{L}^{c_k}(\hat{c}_{i,k}, c_{i,k}) \right\}, \tag{4}$$

where $\alpha > 0$ controls the trade-off between target and concept predictive performance. Observe that parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are optimized simultaneously.

Intervenability. A salient difference between CBMs and multitask models is that a practitioner utilizing a CBM model can interact with it by intervening on concept predictions, e.g. "correcting" the model by setting the predicted values to the ground truth $\hat{c}_{i,k} := c_{i,k}$. In particular,

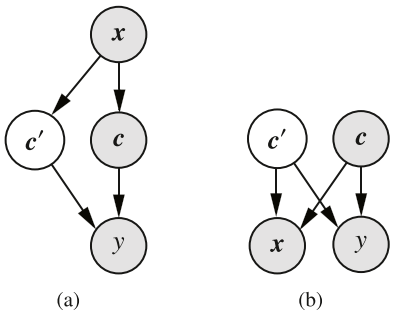

Fig. 3. Generative models that result in incomplete concept sets summarized as directed graphical models. Shaded and unshaded nodes correspond to observed and unobserved variables, respectively. For both (a) and (b), in general, $\boldsymbol{x} \not\perp\!\!\!\perp y \mid \boldsymbol{c}$ since there exists an active path (Geiger et al., 1990) between $\boldsymbol{x}$ and y through unobserved concepts $\boldsymbol{c}'$.

for a data point $1 \leq i \leq N$, the updated prediction after the intervention on the concepts from a subset $S \subseteq \{1, \ldots, K\}$ is given by

$$\hat{y}_i^S = f_{\hat{\theta}}\left(\hat{\boldsymbol{c}}_{\{1,\ldots,K\}\setminus S}, \boldsymbol{c}_S\right), \tag{5}$$

where $\hat{\boldsymbol{c}}$ and $\boldsymbol{c}$ refer to the predicted and ground truth concept vectors, respectively. Note the notation abuse in the order of the arguments in $f_{\hat{\theta}}(\cdot)$.

2.4. Semi-supervised multiview concept bottleneck models

As previously stated, the set of K concepts given at the training may prove incomplete, owing to factors such as the high cost of annotation, the lack of knowledge, or ethical concerns regarding the measurement of certain variables. More formally, concept bottlenecks implicitly assume that concepts are a sufficient statistic for the target variable (Yeh et al., 2020); in other words, $\boldsymbol{x} \perp\!\!\!\perp y \mid \boldsymbol{c}$. A situation where $\boldsymbol{x} \not\perp\!\!\!\perp y \mid \boldsymbol{c}$ may occur when some ground-truth concept variables are systematically missing in the acquired dataset, i.e. unobserved for all data points. Fig. 3 depicts two data-generating mechanisms that may lead to the scenario described above. When this is the case, the predictive performance of the CBM is limited since the model solely relies on the predefined set of concepts which is insufficient. To address this limitation, we propose a semi-supervised variant of the MVCBM (Fig. 1) that additionally learns representations complementary to the concepts and relevant to the downstream prediction task.

Next to the feature extraction and concept prediction, SSMVCBM includes an unsupervised module mapping views $\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}$ to the representation $\hat{\boldsymbol{z}}_i \in \mathbb{R}^J$. To predict the label, $\hat{\boldsymbol{c}}_i$ and $\hat{\boldsymbol{z}}_i$ are concatenated and fed into the target model. This variant of the model is semi-supervised in that the label is predicted based on both $\hat{\boldsymbol{c}}_i$ and $\hat{\boldsymbol{z}}_i$, where $\hat{\boldsymbol{c}}_i$ are supervised by the concept prediction loss (Eq. (7)), while $\hat{\boldsymbol{z}}_i$ are complementary representations learnt without concept labels. Representations $\hat{\boldsymbol{z}}_i$ are meant to capture the residual relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ not represented among the observed concepts $\boldsymbol{c}$. A forward pass of the SSMVCBM is given by

(i) Feature extraction:

$$\begin{aligned} \boldsymbol{h}_i^{c,v} &= \boldsymbol{h}_{\psi^c}\left(\boldsymbol{x}_i^v\right), 1 \leq v \leq V_i, \\ \boldsymbol{h}_i^{z,v} &= \boldsymbol{h}_{\psi^z}\left(\boldsymbol{x}_i^v\right), 1 \leq v \leq V_i, \end{aligned} \tag{6a}$$

(ii) Feature fusion:

$$\begin{aligned} \bar{\boldsymbol{h}}_i^c &= \boldsymbol{r}_{\xi^c}\left(\left\{\boldsymbol{h}_i^{c,v}\right\}_{v=1}^{V_i}\right), \\ \bar{\boldsymbol{h}}_i^z &= \boldsymbol{r}_{\xi^z}\left(\left\{\boldsymbol{h}_i^{z,v}\right\}_{v=1}^{V_i}\right), \end{aligned} \tag{6b}$$

(iii) Concept and representation prediction:

$$\hat{\boldsymbol{c}}_i = \boldsymbol{s}_{\zeta^c}\left(\bar{\boldsymbol{h}}_i^c\right), \tag{6c}$$

$$\hat{\boldsymbol{z}}_i = \boldsymbol{s}_{\zeta^z}\left(\bar{\boldsymbol{h}}_i^z\right),$$

(iv) Label prediction:

$$\hat{y}_i = f_\theta\left(\left[\hat{\boldsymbol{c}}_i, \hat{\boldsymbol{z}}_i\right]\right), \tag{6d}$$

where variables and parameters superscripted by $\boldsymbol{c}$ and $\boldsymbol{z}$ correspond to the concept and representation learning modules, respectively.

To avoid learning a representation redundant to the concepts, it is desirable that $\hat{\boldsymbol{c}} \perp\!\!\!\perp \hat{\boldsymbol{z}} \mid y$, i.e. the predicted concepts and unsupervised representations should be statistically independent conditional on the label. Concretely, we use another neural network $\boldsymbol{a}_{\boldsymbol{\tau}} : \mathbb{R}^J \to \mathbb{R}^K$, parameterized by weights $\boldsymbol{\tau}$, to quantify the degree of statistical dependence as $\max_{\boldsymbol{\tau}} \operatorname{corr}\left(\boldsymbol{a}_{\boldsymbol{\tau}}(\hat{\boldsymbol{z}}), \hat{\boldsymbol{c}}\right)$ (Adeli et al., 2021). Thus, network $\boldsymbol{a}_{\boldsymbol{\tau}}$ is used to adversarially regularize representation $\hat{\boldsymbol{z}}$. Empirically, we observed that this regularization scheme helps de-correlate $\hat{\boldsymbol{z}}$ from concept predictions and improves the model's intervenability (Appendix F.2). Additionally, note that, for the data-generating mechanisms shown in Fig. 3, $\hat{\boldsymbol{z}}$ does not need to identify unobserved concepts $\boldsymbol{c}'$ but rather represents the residual relationship between $\boldsymbol{x}$ and y.

The procedure to train SSMVCBMs is outlined in Algorithm Appendix D.1. Similar to the sequential optimization for (MV)CBMs as in Eqs. (2) and (3), it consists of multiple steps. First, parameters $\boldsymbol{\phi}^c = \{\boldsymbol{\psi}^c, \boldsymbol{\xi}^c, \boldsymbol{\zeta}^c\}$ involved in concept prediction are optimized using the loss function analogous to Eq. (2). Then, we fix $\hat{\boldsymbol{\phi}}^c$ and optimize parameters $\boldsymbol{\phi}^z = \{\boldsymbol{\psi}^z, \boldsymbol{\xi}^z, \boldsymbol{\zeta}^z\}$ by solving the following problem:

$$\begin{aligned} \hat{\boldsymbol{\phi}}^z, \tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\phi}^z, \theta} \max_{\boldsymbol{\tau}} \sum_{i=1}^N w_i^t \mathcal{L}^t\left(\hat{y}_i, y_i\right) - \\ \lambda \sum_{i=1}^N \sum_{k=1}^K w_i^{c_k} \mathcal{L}^{c_k}\left(\left[a_{\boldsymbol{\tau}}\left(\hat{\boldsymbol{z}}_i\right)\right]_k, \hat{c}_{i,k}\right), \end{aligned} \tag{7}$$

where $\lambda > 0$ is a tuning parameter corresponding to the weight of the adversarial regularizer. The loss function above can be extended with further regularization terms, e.g. to de-correlate individual dimensions of $\hat{\boldsymbol{z}}$ (Cogswell et al., 2016), facilitating a more straightforward interpretation. In practice, the minimax objective is optimized using adversarial training similarly to the generative adversarial networks (Goodfellow et al., 2020). Last but not least, parameters of the target model are re-optimized, cf. Eq. (3), treating $\hat{\boldsymbol{\phi}}^c$ and $\hat{\boldsymbol{\phi}}^z$ as fixed: $\hat{\boldsymbol{\theta}} = \arg\min_\theta \sum_{i=1}^N w_i^t \mathcal{L}^t\left(f_\theta\left(\left[\hat{\boldsymbol{c}}_i, \hat{\boldsymbol{z}}_i\right]\right), y_i\right)$.

3. Experiments and results

The purpose of our experiments was twofold: (i) to present a proof of concept for the introduced extensions of the CBMs on simple benchmarks and (ii) to apply our techniques to a real-world medical imaging dataset. In the subsequent sections, we provide a more detailed overview of the experimental setup.

3.1. Experimental setup

Datasets and validation scheme. To test the feasibility of the proposed concept-based multiview classification approaches, we conducted an initial experiment using a synthetic tabular nonlinear classification problem. The generative process of this dataset was defined directly based on the classical concept bottleneck model, involving (i) the sampling of a design matrix, (ii) the mapping of features to concepts, and (iii) the use of these concepts to construct labels. In addition, we constructed multiple "views", each comprising a subset of the original feature set. This dataset is particularly suited to multiview approaches due to its inherent structure. Its essential advantage over the conventional benchmarks from the literature, such as the UCSD Birds, is the presence of reliable per-data-point concept labels. Additional details can be found in Appendix B. This problem features binary concepts that are identifiable from the given multiview observations. Although, herein, concept and target prediction are classification problems, all methods present are easily extendable to regression. In

Table 2
Explanation and descriptive statistics for the concept variables chosen for the pediatric appendicitis dataset. All concept variables are binary. The right-most column reports the percentage of the positive outcome values.

	Name	Description	Pos., %
c_1	Visibility of the appendix	Visibility of the vermiform appendix during the examination	76
c_2	Free intraperitoneal fluid	Free fluids in the abdomen	43
c_3	Appendix layer structure	Characterization of the appendix layers, e.g. irregular in case of an increasing inflammation	14
c_4	Target sign	Axial image of the appendix with the fluid-filled center surrounded by echogenic mucosa and submucosa and hypoechoic muscularis	13
c_5	Surrounding tissue reaction	Inflammation signs in tissue surrounding the appendix	33
c_6	Pathological lymph nodes	Enlarged and inflamed intra-abdominal lymph nodes	21
c_7	Thickening of the bowel wall	Edema of the intestinal wall, >2–3 mm	8
c_8	Coprostasis	Fecal impaction in the colon	6
c_9	Meteorism	Accumulation of gas in the intestine	15

our experiments, we assessed the models' performance at (i) target, (ii) concept prediction, and (iii) the effectiveness of interventions on the predicted concepts. Additionally, to explore the scenario where the set of concepts is incomplete, we purposefully trained the models on concept subsets of varying sizes. We compared the performance of our approach with that of single- and multiview black-box classifiers and the vanilla concept bottlenecks (Koh et al., 2020). In addition to the tabular data, we constructed a semi-synthetic attribute-based natural image dataset based on the *Animals with Attributes 2* (Lampert et al., 2009; Xian et al., 2019) (Appendix C). The experimental results for this benchmark are reported in Appendix F.1.

Last but not least, to demonstrate the effectiveness of our proposed methods on real-world data, we employed ultrasound imaging and tabular clinical, laboratory, and scoring data from pediatric patients with suspected appendicitis. We explored three different target variables encompassing the diagnosis, treatment assignment, and complications. A comprehensive overview of this dataset is available in the previous sections and in Appendix A. For model validation and comparison, we divided the data according to the 90%-10% train-test split. Hyperparameter tuning was performed only on the training set using five-fold cross-validation. The final hyperparameter values are reported in Table Appendix E.2–Appendix E.6. The list of high-level concepts relevant to decision support for pediatric appendicitis can be found in Table 2. The selection criteria for these variables were the following: (*i*) the concept had to be detectable from ultrasound images, as confirmed by a qualified physician, and (*ii*) the variable had to had been collected preoperatively.

Ablations. We compared several variations of the proposed multiview concept bottlenecks to better understand the role of the design choices made. Specifically, we trained models using sequential (MVCBM-seq) and joint (MVCBM-joint) optimization procedures given by Eqs. (2)–(4). We also compared the semi-supervised extension (SSMVCBM) defined in Eq. (6) to the basic MVCBM. To facilitate meaningful comparison, we purposefully trained models under insufficient concept sets to observe if the SSMVCBM could achieve any performance improvement over the MVCBM. Furthermore, we investigated the impact of two fusion functions, namely, the arithmetic mean ((SS)MVCBM-avg) and LSTM ((SS)MVCBM-LSTM). Lastly, similar to Koh et al. (2020), we explored interventions on the concept bottlenecks by replacing the predicted concept values with the ground truth at test time. The goal was to investigate whether a practitioner utilizing a concept-based model could improve its predictions interactively.

Baselines. We benchmarked the performance of the (SS)MVCBMs against several baselines. Across all datasets, we applied single-view neural-network-based classifiers. Specifically, we trained MLPs for tabular data and fine-tuned ResNet-18 (He et al., 2016) on images. As an interpretable single-view baseline, we employed vanilla CBMs. To ensure a fair comparison between CBMs and (SS)MVCBMs, we utilized identical architectures for individual modules. As a black-box multiview baseline, we employed a neural network with the same architecture as for the MVCBM but trained without concept supervision in the bottleneck layer, which we refer to as multiview bottleneck (MVBM). Similarly, as for its interpretable counterpart, we compared two ways of aggregating per-view representations: averaging and LSTM. Lastly, specific to the pediatric appendicitis dataset, in addition to deep-learning- and concept-based approaches, we also investigated an alternative baseline predictive model: a random forest (RF) (Breiman, 2001) fitted on radiomic features (van Griethuysen et al., 2017). The features were extracted from every image and averaged across the views for each subject.

Evaluation. Since the intended use case of our models in healthcare applications is decision support rather than decision-making, we mainly focused on evaluating the performance of concept and label predictions using areas under receiver operating characteristic (AUROC) and precision–recall (AUPR) curves. Notably, for pediatric appendicitis, different metrics may be relevant depending on the target variable, e.g. a low false negative rate may be critical for diagnosis and severity, while a low false positive rate may be desirable for management to avert negative appendectomies (Kryzauskas et al., 2016). Furthermore, for appendicitis, we also assessed the predictions' calibration using the Brier score.

Implementation details. We implemented MVCBM and SSMVCBM in PyTorch (v 1.11.0) (Paszke et al., 2019). Across all experiments and models, when applicable, we fine-tuned pretrained ResNet-18 (He et al., 2016) as the shared view encoder. For the concept encoder and target model, we utilized MLPs with ReLU hidden activations. Detailed architecture specifications are provided in Appendix E.

We used the *PyRadiomics* package (van Griethuysen et al., 2017) for radiomic feature extraction. Features were extracted from the whole images without prior segmentation of the region of interest since segmentation is beyond the scope of the current work. We computed first-order statistics, gray level size zone and gray level run length matrix features from the original and square-filtered images. Random forests were trained with a cost-sensitive loss function to account for class imbalance. ANOVA F-value-based feature selection was performed using nested cross-validation to improve the performance of this baseline further. The remainder of the implementation details can be found in Appendix E and within the publicly available code and documentation.

3.2. Proof of concept on synthetic data

The first benchmark we considered was tabular synthetic nonlinear data. Fig. 4 contains the summary of the results. As expected, black-box and concept-based multiview approaches are consistently more accurate than their single-view counterparts at target (Fig. 4(a)) and concept prediction (Fig. 4(b)). Namely, a multiview bottleneck model without concept supervision (MVBM) performs considerably better than a multilayer perceptron trained on a single view (MLP) (paired t-test p-value < 0.0001 for target AUROC); similarly, a multiview concept bottleneck (MVCBM) outperforms a simple CBM (for all numbers of

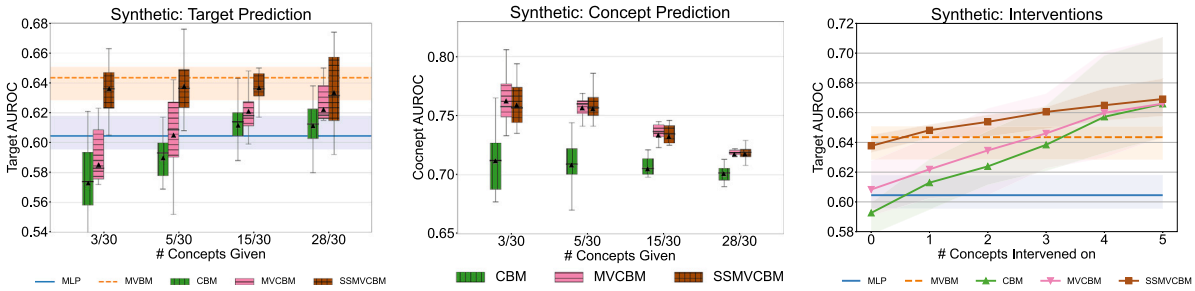

Fig. 4. Target and concept prediction results on synthetic data for the proposed multiview concept bottleneck (MVCBM) and semi-supervised multiview concept bottleneck (SSMVCBM) models alongside several baselines. All plots were produced across ten independent simulations. (a) One-vs-all AUROCs for predicting the target on the test data under the varying number of observed concepts. MLP and MVBM do not rely on concepts; their AUROCs are shown as horizontal lines for reference. (b) AUROCs for predicting concepts on the test data under the varying number of observed concepts. AUROCs were averaged across the observed concepts. (c) AUROCs for predicting the target on the test data after intervening on the varying number of concepts. The intervention experiment was performed for 5/30 observed concepts, i.e. under an incomplete concept set. The performances of non-intervenable MLP and MVBM baselines are shown as horizontal lines. Confidence bands correspond to interquartile ranges across independent simulations and several randomly sampled concept subsets.

concepts given, p-value < 0.05 for target and concept AUROC). Notably, the target prediction accuracy for CBM and MVCBM increases with the number of concepts given, as shown in Fig. 4(a). When almost a complete concept set is provided, the performance of the multiview CBM becomes closer to that of the multiview black-box classifier. The semi-supervised MVCBM (SSMVCBM) performs well even when very few concepts are known and is close to the black-box baseline in most settings (for at least 5/30 concepts given, p-value > 0.05 for target AUROC).

For the concept prediction, MVCBM and SSMVCBM attain comparable performance with higher AUROCs than the single-view model (Fig. 4(b)). As expected, the semi-supervised model predicts the concepts equally well compared to the MVCBM (for all numbers of concepts given, p-value > 0.05 for concept AUROC); thus, representation learning has no effect on the concept prediction. Lastly, we observe from Fig. 4(c) that similarly to the classical CBM, both multiview variants are intervenable, i.e. their predictive performance improves when replacing predicted concepts with the ground truth at test time.

In addition to the results above, Appendix F.1 describes experiments on a semi-synthetic attribute-based natural image dataset. In brief, we observed similar results to the ones reported in Fig. 4. In Appendix F.2, we explore the SSMVCBM in more detail, performing an ablation study on the effect of adversarial regularization.

3.3. Application to pediatric appendicitis

Our multiview concept bottleneck models are readily applicable to medical imaging datasets, which, in practice, often include multiple views and heterogeneous data types. In the following, we explore the application of the multiview CBMs to the pediatric appendicitis dataset.

Predicting high-level ultrasound features. We first evaluated the ability of all concept-based models to predict high-level appendix ultrasound features (Table 2) from (multiple) abdominal US images. Table 3 contains test-set AUROCs and AUPRs achieved by the different variants of the concept bottleneck. In addition to comparing vanilla CBMs to their multiview and semi-supervised extensions, we investigated the effect of the optimization procedure, sequential vs. joint, and view-specific feature fusion, averaging vs. long short-term memory (LSTM). The models included in Table 3 were trained to predict the diagnosis (*appendicitis* vs. *no appendicitis*); however, we observed similar results for the management and severity, as shown in Tables 4–5. Minor discrepancies across the three classification problems are attributable to the differences in the weights assigned to data points in the cost-sensitive loss function (Eqs. (2)–(4) and (7)) and the choice of hyperparameter values (Table Appendix E.2–Appendix E.6).

Across all target variables, most concepts could be predicted by at least one of the models significantly better than by a fair coin flip (one-sample two-sided t-test p-value < 0.05, adjusted using the Benjamini–Yekutieli procedure with the FDR of $q = 0.05$). Surprisingly, some of the variables with relatively few cases present in the dataset could be captured by some models, e.g. *coprostasis* (c_8) and *meteorism* (c_9) by the LSTM-based variants of MVCBM and SSMVCBM. On the other hand, the *thickening of the bowel wall* (c_7) was particularly challenging to model, likely due to its low prevalence and the lack of predictive power in the downstream classification task: some models trained with the severity as the target were able to perform significantly better than random, as shown in Table 5.

Note that, in a few cases, some models achieved average AUROCs below the expected performance of a fair coin flip (Table 3), e.g. both sequentially and jointly optimized CBMs attained an AUROC close to 0.40 for predicting *meteorism*. Such performance is attributable to the sparsity of some concept variables; for instance, only 15% of subjects had a positive label for meteorism (Table 2). Another factor is the use of weighted loss functions for the concept and target prediction (Eqs. (2)–(4) and (7)). Consequently, the models may over-predict the minority class and perform worse than a fair coin flip.

Predictably, sequentially optimized models (seq) were more performant at the concept prediction than the ones optimized jointly (joint), in agreement with the findings reported in the literature (Koh et al., 2020). Similar to the experiments on the synthetic data shown in Fig. 4(b), the models aggregating multiple views tended to have higher AUROCs and AUPRs. However, by contrast, LSTM-based aggregation consistently and noticeably outperformed simple averaging (avg), especially for predicting the visibility of the appendix—one of the most important diagnostic concepts (Marcinkevics et al., 2021). This could be associated with the loose spatiotemporal ordering among the US images acquired for each subject. Last but not least, semi-supervised bottlenecks were comparable to the sequentially optimized MVCBMs. Thus, learning complementary representations disentangled from the concepts did not hurt the model's performance at concept prediction.

In addition to the discriminative power, we assessed the calibration of the concept predictions. The test-set Brier scores across the three targets are reported in Appendix F.5, Table Appendix F.4. Overall, similar to the findings above, multiview models attained lower Brier scores for most concept variables than the single-view CBMs. The cases wherein single-view CBMs performed better than their multiview counterparts may be attributed to the imbalances in concept distributions and the fact that the Brier score does not adjust for such situations. For instance, for very sparse response variables, a classifier trivially predicting the most frequent category would achieve a relatively low Brier score. Although many models predicted several concepts significantly better than the constant prediction of 0.5, their Brier scores were mainly in the range of 0.18-0.23, which is not considerably below the baseline of 0.25.

Table 3
Models' test-set performance at concept prediction on the pediatric appendicitis dataset with the diagnosis as the target variable. Test-set AUROCs and AUPRs are reported as averages and standard deviations across ten independent initializations. Herein, "seq" and "joint" denote sequential and joint optimization, respectively, whereas "avg" and "LSTM" stand for the averaging- and LSTM-based fusion. Bold indicates the best result; italics indicates the second best. The meaning of the concept variables: c_1, visibility of the appendix; c_2, free intraperitoneal fluid; c_3, appendix layer structure; c_4, target sign; c_5, surrounding tissue reaction; c_6, pathological lymph nodes; c_7, thickening of the bowel wall; c_8, coprostasis; c_9, meteorism.

Metric	Model	Concept								
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
AUROC	Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	CBM-seq	0.52±0.04	0.47±0.04	0.60±0.07*	0.56±0.08	0.63±0.05*	0.57±0.05*	0.45±0.08	0.48±0.08	0.39±0.07
	CBM-joint	0.50±0.05	0.47±0.03	0.57±0.05*	0.54±0.06	0.64±0.04*	0.59±0.05*	0.39±0.06	0.57±0.12	0.38±0.09
	MVCBM-seq-avg	0.61±0.05*	0.49±0.05	0.66±0.08*	0.60±0.08*	0.51±0.08	0.66±0.08*	*0.50±0.04*	0.47±0.12	0.55±0.07
	MVCBM-seq-LSTM	*0.83±0.03**	*0.59±0.03**	0.62±0.04*	**0.71±0.04***	0.65±0.04*	0.67±0.07*	0.49±0.07	**0.68±0.10***	*0.73±0.06**
	MVCBM-joint-avg	0.55±0.10	0.47±0.07	**0.73±0.07***	0.63±0.07*	0.61±0.06*	0.63±0.07*	0.48±0.06	0.45±0.13	0.54±0.11
	MVCBM-joint-LSTM	**0.85±0.03***	0.55±0.04*	0.58±0.04*	*0.70±0.03**	**0.75±0.02***	0.55±0.09	0.45±0.12	0.68±0.17	**0.77±0.03***
	SSMVCBM-avg	0.62±0.05*	**0.60±0.05***	*0.72±0.05**	0.67±0.05*	0.54±0.05	*0.68±0.08**	**0.53±0.11**	0.43±0.08	0.47±0.07
	SSMVCBM-LSTM	**0.85±0.04***	0.58±0.06*	0.66±0.05*	**0.71±0.06***	*0.67±0.04**	**0.69±0.06***	0.45±0.09	*0.66±0.11**	*0.73±0.05**
AUPR	Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
	CBM-seq	0.71±0.03	0.53±0.03*	0.29±0.06*	0.26±0.05	0.64±0.05*	0.38±0.06*	0.15±0.03	0.12±0.02	0.11±0.02
	CBM-joint	0.73±0.05	0.49±0.04	0.30±0.06*	0.30±0.08	0.64±0.05*	0.38±0.09*	0.15±0.05	0.19±0.08	0.11±0.02
	MVCBM-seq-avg	0.79±0.04*	0.53±0.06	*0.34±0.10**	0.35±0.10*	0.53±0.07	*0.41±0.07**	0.17±0.04	0.14±0.04	0.25±0.12
	MVCBM-seq-LSTM	0.92±0.02*	*0.59±0.04**	0.32±0.05	**0.38±0.04***	*0.67±0.04**	**0.42±0.10***	0.15±0.02	*0.21±0.08*	**0.40±0.11***
	MVCBM-joint-avg	0.75±0.08	0.48±0.06	**0.38±0.09***	0.30±0.06	0.58±0.05*	0.39±0.08*	**0.21±0.08**	0.15±0.08	0.16±0.05
	MVCBM-joint-LSTM	**0.94±0.01***	0.50±0.05	0.26±0.08	*0.37±0.07**	**0.74±0.04***	0.32±0.09	0.16±0.08	**0.31±0.20**	0.28±0.07*
	SSMVCBM-avg	0.79±0.04*	0.58±0.03*	**0.38±0.05***	0.34±0.04*	0.54±0.06	**0.42±0.08***	*0.20±0.06*	0.12±0.04	0.17±0.07
	SSMVCBM-LSTM	*0.93±0.03**	**0.60±0.06***	0.31±0.06*	**0.38±0.06***	*0.67±0.04**	0.39±0.06*	0.19±0.06	0.19±0.07	*0.30±0.09**

* AUROCs and AUPRs that are significantly greater than the expected performance of a fair coin flip (random).

Table 4
Models' test-set performance at concept prediction on the appendicitis dataset with the management as the target variable. Test-set AUROCs and AUPRs are reported as averages and standard deviations across ten independent initializations. Herein, "seq" and "joint" denote sequential and joint optimization, respectively, whereas "avg" and "LSTM" stand for the averaging- and LSTM-based fusion. Bold indicates the best result; italics indicates the second best. The meaning of the concept variables: c_1, visibility of the appendix; c_2, free intraperitoneal fluid; c_3, appendix layer structure; c_4, target sign; c_5, surrounding tissue reaction; c_6, pathological lymph nodes; c_7, thickening of the bowel wall; c_8, coprostasis; c_9, meteorism.

Metric	Model	Concept								
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
AUROC	Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	CBM-seq	0.51±0.05	0.54±0.07	0.63±0.05*	0.49±0.07	0.65±0.07*	0.56±0.06	0.47±0.10	0.60±0.10	0.54±0.07
	CBM-joint	0.54±0.08	0.51±0.08	0.64±0.06*	0.49±0.06	*0.67±0.03**	0.54±0.07	0.49±0.07	0.56±0.10	0.47±0.09
	MVCBM-seq-avg	0.62±0.06*	0.48±0.07	0.69±0.03*	0.54±0.12	0.49±0.08	0.60±0.07*	0.48±0.09	0.47±0.13	0.57±0.09
	MVCBM-seq-LSTM	**0.86±0.05***	*0.55±0.05*	0.62±0.05*	*0.69±0.03**	0.66±0.04*	**0.65±0.06***	*0.50±0.07*	**0.75±0.09***	**0.74±0.06***
	MVCBM-joint-avg	0.52±0.07	0.53±0.06	*0.71±0.07**	0.59±0.05*	0.64±0.07*	**0.65±0.04***	0.48±0.10	0.54±0.07	0.52±0.15
	MVCBM-joint-LSTM	0.80±0.05*	0.41±0.08	0.66±0.07*	0.61±0.04*	0.66±0.03*	*0.62±0.07**	**0.51±0.07**	0.62±0.11	0.63±0.08*
	SSMVCBM-avg	0.62±0.07*	**0.57±0.08**	**0.73±0.04***	0.63±0.05*	0.55±0.04	**0.65±0.07***	*0.50±0.08*	0.49±0.08	0.52±0.05
	SSMVCBM-LSTM	*0.84±0.02**	0.54±0.05	0.70±0.05*	**0.70±0.03***	**0.68±0.05***	*0.62±0.07**	*0.50±0.10*	*0.72±0.05**	*0.72±0.10**
AUPR	Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
	CBM-seq	0.76±0.03	*0.55±0.07*	0.37±0.09*	0.23±0.03	*0.66±0.07**	0.35±0.10	*0.19±0.06*	0.20±0.13	0.17±0.03
	CBM-joint	0.77±0.04*	0.51±0.06	**0.45±0.08***	0.24±0.07	0.64±0.04*	0.29±0.04	*0.19±0.05*	0.17±0.09	0.15±0.06
	MVCBM-seq-avg	0.79±0.04*	0.52±0.08	0.35±0.04*	0.31±0.14	0.51±0.06	0.37±0.08*	0.17±0.04	0.12±0.04	0.18±0.05
	MVCBM-seq-LSTM	**0.95±0.02***	*0.55±0.03**	0.32±0.08*	**0.38±0.04***	*0.66±0.03**	*0.38±0.09**	0.16±0.02	**0.30±0.16**	**0.30±0.06***
	MVCBM-joint-avg	0.71±0.04	0.53±0.05	0.36±0.10*	0.28±0.03*	0.60±0.07*	**0.39±0.06***	0.17±0.05	0.20±0.07	0.21±0.10
	MVCBM-joint-LSTM	0.91±0.03*	0.44±0.05	0.31±0.06*	0.33±0.06*	0.64±0.03*	*0.38±0.06**	*0.19±0.04*	0.19±0.11	*0.28±0.14*
	SSMVCBM-avg	0.78±0.06	**0.60±0.07***	*0.41±0.08**	0.33±0.08*	0.55±0.05	**0.39±0.07***	**0.22±0.06**	0.12±0.02	0.23±0.08
	SSMVCBM-LSTM	*0.93±0.01**	*0.55±0.06*	0.38±0.09*	*0.37±0.06**	**0.67±0.06***	0.35±0.06*	0.17±0.05	*0.24±0.05**	0.27±0.08*

* AUROCs and AUPRs that are significantly greater than the expected performance of a fair coin flip (random).

Predicting diagnosis, management, and severity. As mentioned, the end goal of the developed models was the prediction of the (i) diagnosis, (ii) management, and (iii) severity among suspected appendicitis patients based on the multiview US images. Test-set performance for these three target variables is reported in Table 6.

With respect to AUROC and AUPR, all models were able to predict all target variables better than the naive baseline. Among the concept-based approaches, multiview models offered a consistent improvement over the vanilla CBM for diagnosis and severity. Moreover, the best-performing concept-based classifiers often achieved AUROCs and AUPRs comparable to those of the black-box MVBM. For the diagnosis, on average, multiview concept bottlenecks with the LSTM-based fusion outperformed averaging-based approaches. However, for management, the opposite was true. Expectedly, while the LSTM-based fusion was helpful in the pediatric appendicitis dataset where US images are chronologically ordered, at test time, the target prediction performance of the LSTM-based CBMs was sensitive to the order of input images, as observed in the supplementary experimental results in Appendix F.3. For the diagnosis and management prediction, we also observed that neural-network-based methods, overall, outperformed

Table 5
Models' test-set performance at concept prediction on the appendicitis dataset with the severity as the target variable. Test-set AUROCs and AUPRs are reported as averages and standard deviations across ten independent initializations. Herein, "seq" and "joint" denote sequential and joint optimization, respectively, whereas "avg" and "LSTM" stand for the averaging- and LSTM-based fusion. Bold indicates the best result; italics indicates the second best. The meaning of the concept variables: c_1, visibility of the appendix; c_2, free intraperitoneal fluid; c_3, appendix layer structure; c_4, target sign; c_5, surrounding tissue reaction; c_6, pathological lymph nodes; c_7, thickening of the bowel wall; c_8, coprostasis; c_9, meteorism.

Metric	Model	Concept								
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
AUROC	Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	CBM-seq	0.51±0.04	*0.58±0.06**	0.61±0.08*	0.52±0.09	0.62±0.04*	0.62±0.05*	0.47±0.09	0.57±0.11	0.50±0.08
	CBM-joint	0.55±0.06	0.46±0.06	0.66±0.06*	0.47±0.06	*0.64±0.04**	0.53±0.07	0.50±0.07	0.58±0.10*	0.49±0.04
	MVCBM-seq-avg	0.54±0.08	0.55±0.04	**0.72±0.07***	0.62±0.04*	0.50±0.05	0.64±0.06*	0.51±0.10	0.47±0.11	0.54±0.10
	MVCBM-seq-LSTM	**0.82±0.04***	0.53±0.04	0.62±0.04*	**0.69±0.04***	0.62±0.05*	**0.72±0.05***	**0.64±0.06***	**0.78±0.03***	**0.70±0.06***
	MVCBM-joint-avg	0.54±0.09	0.51±0.06	0.70±0.06*	0.59±0.08*	0.61±0.06*	0.62±0.05*	0.54±0.15	0.48±0.14	0.55±0.12
	MVCBM-joint-LSTM	**0.82±0.03***	0.48±0.06	0.66±0.07*	0.64±0.06*	**0.65±0.05***	0.64±0.09*	0.47±0.09	0.61±0.14	*0.65±0.05**
	SSMVCBM-avg	0.53±0.06*	0.56±0.08*	*0.71±0.05**	0.60±0.06*	0.51±0.05	0.64±0.09*	0.46±0.08	0.48±0.09	0.53±0.03
	SSMVCBM-LSTM	*0.77±0.10**	**0.59±0.08**	0.70±0.06*	*0.67±0.07**	**0.65±0.07***	*0.67±0.05**	*0.62±0.08**	*0.74±0.15**	0.64±0.11*
AUPR	Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
	CBM-seq	0.75±0.03	*0.58±0.05**	0.34±0.09*	0.24±0.05	0.64±0.04*	0.35±0.06*	0.18±0.05	0.19±0.07	0.15±0.03
	CBM-joint	0.77±0.05	0.47±0.04	*0.37±0.09**	0.25±0.06	0.64±0.05*	0.30±0.07	0.17±0.04	0.18±0.06	0.18±0.08
	MVCBM-seq-avg	0.75±0.05	*0.58±0.06**	**0.42±0.07***	0.33±0.06*	0.53±0.05	0.41±0.08*	0.21±0.05	0.13±0.05	*0.24±0.12*
	MVCBM-seq-LSTM	*0.91±0.04**	0.55±0.04*	0.33±0.08*	**0.40±0.06***	*0.65±0.03**	**0.50±0.11***	*0.23±0.05**	*0.27±0.05**	**0.26±0.07***
	MVCBM-joint-avg	0.74±0.06	0.51±0.07	**0.42±0.09***	0.28±0.07	0.59±0.06*	0.35±0.05*	0.22±0.06	0.22±0.13	0.21±0.08
	MVCBM-joint-LSTM	**0.92±0.02***	0.49±0.05	*0.37±0.11**	0.32±0.07*	*0.65±0.06**	0.39±0.07*	0.20±0.06	0.17±0.07	0.21±0.06*
	SSMVCBM-avg	0.73±0.05	*0.58±0.07**	0.36±0.05*	0.28±0.04*	0.53±0.05	0.37±0.09*	0.20±0.06	0.13±0.02	*0.24±0.06**
	SSMVCBM-LSTM	0.88±0.06*	**0.60±0.06***	**0.42±0.06***	*0.39±0.09**	**0.67±0.07***	*0.43±0.10**	**0.24±0.08**	**0.30±0.13***	0.20±0.05*

* AUROCs and AUPRs that are significantly greater than the expected performance of a fair coin flip (random).

Table 6
Models' test-set performance at predicting diagnosis, management, and severity. Test-set AUROCs, AUPRs, and Brier scores are reported as averages and standard deviations across ten independent initializations. Bold indicates the best result; italics indicates the second best.

Model	Diagnosis			Management			Severity		
	AUROC	AUPR	Brier	AUROC	AUPR	Brier	AUROC	AUPR	Brier
Random	0.50	0.75	0.25	0.50	0.47	0.25	0.50	0.23	0.25
Radiomics + RF	0.64±0.02	0.82±0.01	0.22±0.00	0.65±0.01	0.60±0.02	*0.24±0.00*	*0.77±0.02*	*0.58±0.04*	**0.15±0.00**
ResNet-18	0.70±0.07	0.88±0.04	0.25±0.08	0.69±0.07	*0.71±0.08*	0.27±0.05	0.73±0.10	0.52±0.10	0.18±0.04
CBM-seq	0.64±0.06	0.84±0.04	0.22±0.02	0.68±0.05	0.68±0.05	**0.23±0.02**	0.66±0.06	0.41±0.08	0.23±0.04
CBM-joint	0.62±0.04	0.83±0.04	0.24±0.02	0.66±0.06	0.68±0.04	**0.23±0.02**	0.68±0.06	0.44±0.08	0.23±0.02
MVBM-avg	*0.76±0.05*	0.89±0.04	0.22±0.03	0.71±0.04	0.69±0.04	0.24±0.02	0.71±0.12	0.59±0.11	0.20±0.05
MVBM-LSTM	*0.76±0.04*	*0.91±0.02*	0.23±0.02	0.67±0.04	0.61±0.04	**0.23±0.02**	0.74±0.13	*0.58±0.12*	0.22±0.07
MVCBM-seq-avg	0.67±0.05	0.85±0.05	0.23±0.02	0.58±0.05	0.62±0.06	0.26±0.02	0.75±0.07	0.56±0.12	0.23±0.04
MVCBM-seq-LSTM	0.73±0.03	0.89±0.01	0.24±0.04	0.57±0.03	0.53±0.04	0.26±0.01	0.70±0.11	0.48±0.16	0.21±0.03
MVCBM-joint-avg	0.66±0.09	0.84±0.06	0.24±0.06	0.69±0.06	0.66±0.11	**0.23±0.02**	0.70±0.06	0.53±0.11	0.24±0.02
MVCBM-joint-LSTM	0.72±0.02	0.88±0.02	0.22±0.01	0.57±0.05	0.50±0.04	0.26±0.01	0.65±0.07	0.37±0.10	0.24±0.02
SSMVCBM-avg	**0.80±0.03**	**0.92±0.02**	*0.20±0.03*	**0.72±0.05**	**0.72±0.04**	0.27±0.05	0.73±0.07	0.57±0.09	*0.17±0.02*
SSMVCBM-LSTM	**0.80±0.06**	**0.92±0.04**	**0.19±0.04**	0.70±0.03	0.67±0.06	0.27±0.04	**0.78±0.05**	*0.58±0.10*	0.21±0.10

RFs fitted on radiomics features. The latter result is not surprising, given that we did not utilize manually segmented regions of interest for radiomics feature extraction. Lastly, across all targets, the semi-supervised extension of the MVCBM achieved higher AUROCs and AUPRs or was comparable to the approaches that purely relied on the concepts.

Brier score results partially agree with AUROCs and AUPRs; however, they feature less variability across model classes. For all target variables, most scores are ≥ 0.20. Combined with the reported AUROCs and AUPRs, the latter finding indicates that the probabilistic predictions of the models considered could benefit from calibration, which could help produce more interpretable probabilistic outputs.

Along with the model comparison w.r.t. AUROCs, AUPRs, and Brier scores, we investigated the tradeoff between true positive (TPR) and false positive (FPR) rates in more detail for predicting the diagnosis. Full results are reported in Table Appendix F.3 (Appendix F.4). In particular, we assessed the models' FPRs for a few fixed satisfactory levels of the TPR. As expected, we observed that, for all approaches, attaining high TPRs led to relatively high FPRs of $> 30\%$.

In summary, concept-based classification on multiview US data is encouragingly effective at predicting the diagnosis. For management, aggregating multiple US images offers no improvement over simple single-view classification. We attribute this to the *diagnostic* nature of the chosen concepts and their limited predictive power for the treatment assignment. Likewise, accurately predicting appendicitis severity is challenging, likely, due to the low prevalence of complicated appendicitis cases in the current dataset. Last but not least, in all tasks, the proposed SSMVCBM mitigated the poorer discriminative performance of concept-based approaches by learning representations complementary to the probably incomplete concept set.

Interacting with the model. The practical utility of CBMs lies in the ability of the human user, in the current use case, the physician, to intervene on the concepts predicted by the model, thus affecting the model's behavior at test time. Similarly to the proof-of-concept experiments, we intervened on the bottleneck layers of the CBM, MVCBM, and SSMVCBM trained on the pediatric appendicitis data. Fig. 5 summarizes these results. Since LSTM-based and sequentially trained classifiers generally captured the concepts better (Table 3), we only considered

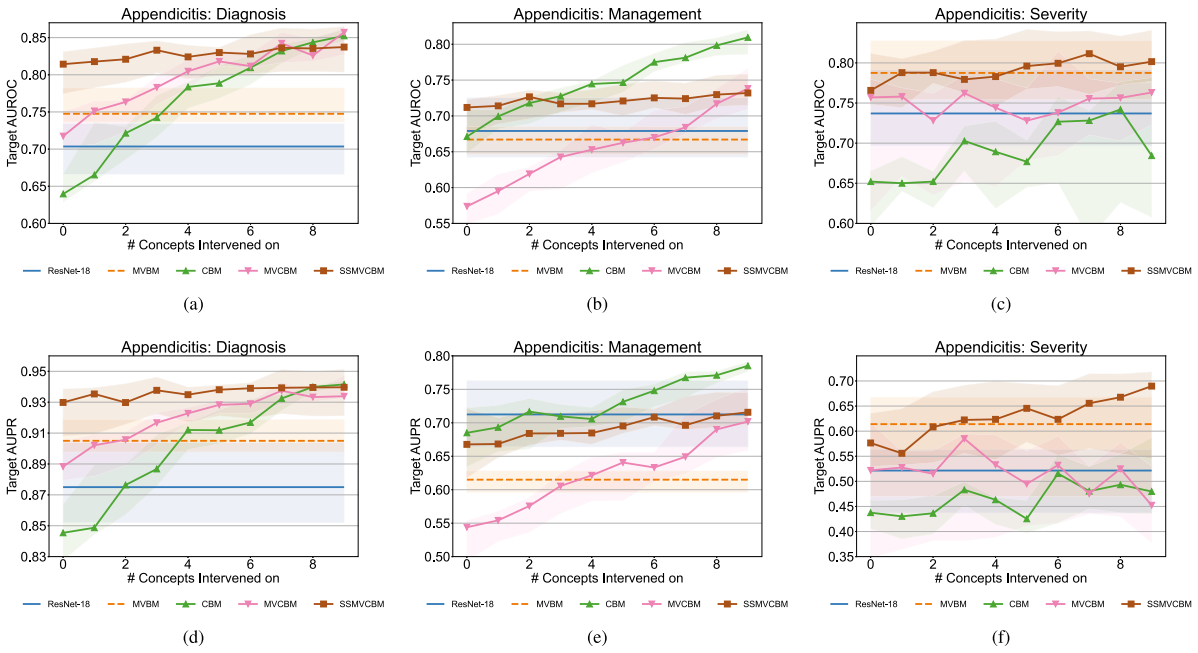

Fig. 5. Intervention experiment results for the pediatric appendicitis dataset. Interventions were performed by replacing the concept values predicted by the concept-based models with the ground truth for the (a,d) diagnosis, (b,e) management, and (c,f) severity as target variables. Lines correspond to median (a–c) test-set AUROCs and (d–f) AUPRs attained by intervened models across ten initializations and three randomly sampled concept subsets. The performances of non-intervenable ResNet-18 and MVBM baselines are shown as horizontal lines.

this specific configuration. Fig. 5 shows the effect of interventions on the three models for the diagnosis (Fig. 5(a) and (d)), management (Fig. 5(b) and (e)), and severity (Fig. 5(c) and (f)). The lines show changes in AUROCs and AUPRs when intervening on randomly chosen concept subsets of varying sizes.

For the diagnosis, the intervention effect is similar to the behavior of the models on the synthetic data shown in Fig. 4(c). Namely, AUROC and AUPR increase steadily with the number of concepts intervened on: for all models, the maximum median AUROC and AUPR achieved are approx. 0.85 and 0.94, respectively. Being the best-performing model (Table 6), SSMVCBM demonstrates only a slight increase in median predictive performance after intervening on the full concept set.

Similarly, for management, we observed an increase in AUROC and AUPR. However, for predicting this target, a single-view CBM performed surprisingly well and overtook multiview models after interventions. Last but not least, interventions yielded no visible performance improvement for severity, possibly, due to considerable variance across initializations and randomly sampled concept subsets.

3.4. Online prediction tool

As a first step towards enabling clinicians and other interested parties to benefit from ML-based decision support, we developed and published an online decision support tool based on the abovementioned methods, available at https://papt.inf.ethz.ch/mvcbm. The use case is illustrated in Fig. 6. The tool utilizes the multiview CBM model (Fig. 1) for predicting the diagnosis in suspected appendicitis patients. The user may upload several ultrasonography images, each representing a different view of the same patient. Image preprocessing, described in the *Methods* section and demonstrated in Fig. 2, may be optionally executed. In addition to predicting the diagnosis, the tool allows the user to intervene on the concept predictions (Table 2) by setting corresponding sigmoid activations to 0 (negative) or 1 (positive). Uploaded images are protected using server-side sessions, which are only temporarily stored on the server and are purged after 30 min. See Appendix G for more information.

4. Discussion

Most of the prior work on using ML for appendicitis has focused on tabular datasets with handcrafted features (Hsieh et al., 2011; Deleger et al., 2013; Reismann et al., 2019; Aydin et al., 2020; Akmese et al., 2020; Stiel et al., 2020; Marcinkevics et al., 2021; Roig Aparicio et al., 2021; Xia et al., 2022) or more invasive imaging modalities, such as computed tomography (Rajpurkar et al., 2020). This work takes the first step towards the computer-aided diagnosis of appendicitis based on abdominal ultrasound, a noninvasive, accessible, and cheap technique. Moreover, to facilitate the replication of our results and allow for comparison with new methods, we made our anonymized dataset publicly accessible. It includes laboratory, physical exam, clinical, and US data from 579 patients. In addition, for demonstratory purposes, we deployed the MVCBM model for the diagnosis as an easy- and free-to-use web tool.

Although appendicitis is a common condition in the pediatric population, diagnosing it and choosing the best therapeutic option is challenging. Early differentiation between simple and complicated, necrotizing appendicitis is crucial for effective management and prognosis (Reddan et al., 2016; Reismann et al., 2019; Kiss et al., 2021). The advances in US resolution, especially with the high-frequency sonography, support the detection of a normal appendix and the identification of indirect appendicitis signs, such as surrounding tissue inflammation and the reaction of the intestinal bowel wall (Park et al., 2011). ML-based decision support tools may further increase diagnostic accuracy and prove pivotal in improving treatment outcomes. The results of the current study are promising, as they suggest that direct interpretation of US images by ML models is a feasible goal. Predictive models, such as the ones developed in this study, may assist physicians in interpreting acquired US images and may even enable comparison of the results with the newly conducted US exams to characterize the progress or resolution of the inflammation.

Moreover, this work presents an improvement upon traditional concept bottleneck models (Koh et al., 2020), making them more readily applicable to medical imaging datasets where multiple images

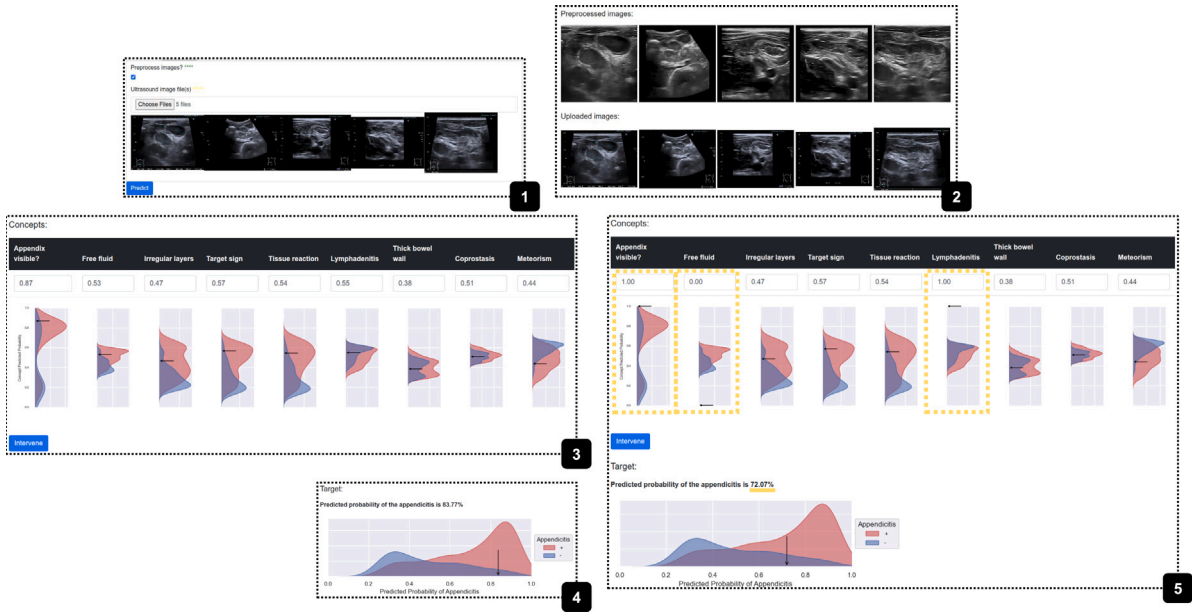

Fig. 6. An illustrated use case for the pediatric appendicitis online prediction tool. (1) The user uploads input ultrasound images corresponding to a single patient. (2) Optionally, preprocessing is performed, and the tool displays original and preprocessed US images. (3) The tool displays predicted concept values, given by sigmoid activations from the corresponding units in the concept bottleneck layer, alongside predicted value histograms obtained from the *training* data (plotted separately for appendicitis and non-appendicitis cases in red and blue, respectively). The user can compare current concept predictions to those made for labeled *training* data points. (4) The tool shows the prediction for the target variable, i.e. for the diagnosis, given by the sigmoid activation. (5) The user may choose to intervene on the concept predictions and, thus, affect the target prediction. For example, if the user was trained in interpreting ultrasound images, they may correct wrongly predicted concepts by setting corresponding variables to 0 (negative) or 1 (positive). In this example, concept predictions that were intervened on (chosen arbitrarily for demonstration) are indicated by yellow dotted lines.

or modalities may be observable for each subject. In order to accomplish this, we proposed a practical architecture based on the hybrid fusion approach (Baltrušaitis et al., 2019), which can effectively handle varying numbers of views per data point, partial observability of the concepts from individual images, and the incorporation of spatial or temporal ordering. While prior research has explored the use of averaging and LSTM techniques for aggregating representations (Havaei et al., 2016; Ma et al., 2019), our focus is specifically on interpretable models, particularly those involving concept-based classification. To the best of our knowledge, this problem setting has not been previously discussed in the literature despite its relevance to biomedical applications (Wang et al., 2020; Qian et al., 2021).

Another scenario that we studied, similarly pertinent to applications, is when the concept set given to a CBM is insufficient (Yeh et al., 2020), i.e. does not entirely capture the predictive relationship between the covariates and the target. To address this issue and improve the CBM's predictive performance, our model learns additional representations complementary to the concepts, i.e. de-correlated from the concepts yet helpful in the downstream prediction problem. To achieve this objective, we modified the model's architecture, incorporated an adversarial regularization term into the loss function, and adapted the training procedure accordingly.

A few previous works have investigated related limitations of the CBMs when the concept set provided to the CBM proves insufficient, and have explored alternative model designs. For instance, Sawada and Nakamura (2022) combined CBMs with self-explaining neural networks to learn additional unsupervised concepts; however, they did not investigate the disentanglement of the given and learned concepts or the intervenability of their extended bottleneck layer. Yuksekgonul et al. (2022) proposed fitting a concept bottleneck *post hoc* for a pretrained backbone and utilized residual fitting to compensate for an incomplete concept set. Moreover, they investigated the global model edition, e.g. to mitigate the classifier's reliance on spurious correlation. In contrast, our work assumes an *ante hoc* modeling scenario and focuses on the local, i.e. single-data-point, interventions. Another related line of research also studied the problem of unobserved concepts and concept leakage (Marconato et al., 2022), employing generative representation learning, which may be challenging to apply to smaller datasets in practice. The most closely related is the concurrent work by Havasi et al. (2022), who extended the standard CBM architecture with a side channel to learn latent concepts and compensate for insufficiency. While their method is similar to ours, it does not address multiview learning or consider medical imaging data.

In our experiments, we have demonstrated the feasibility of the proposed models and the benefits of the multiview and semi-supervised concept-based approach on synthetic and medical image data. Our findings have shown that the MVCBM and SSMVCBM models have generally outperformed vanilla CBM in terms of both concept and target prediction. Moreover, based on the US data, we have developed predictive models for appendicitis, its severity and the management of pediatric patients with abdominal pain (Tables 3–6). Our results suggest that, for the diagnosis, multiview concept bottlenecks can achieve comparable performance to black-box models while allowing medical practitioners to interpret and intervene on the predictions. For management and severity, we observed somewhat inconclusive results with little difference across the single- and multiview classifiers. We attribute the latter to the limited predictive power of the ultrasonographic findings for these targets (Marcinkevics et al., 2021), the diagnostic nature of the chosen concepts and the overall moderate size of the training set. For instance, it had been previously shown that the most important predictor of the treatment assignment is peritonitis/abdominal guarding (Marcinkevics et al., 2021) assessed during a clinical examination. Among the US findings, most other predictively useful attributes can be identified based on the RLQ image alone. Therefore, we hypothesize that the additional views, e.g. depicting pathological lymph nodes or meteorism, are not as helpful for the management classification. This observation might explain the relatively worse performance of the multiview approaches for this target variable.

Nevertheless, the current study exhibits certain limitations with regard to its design, experimental setup, and proposed methods. The appendicitis dataset represents a moderately-sized and relatively homogeneous patient cohort recruited from a single clinical center over a

short time (between 2016 and 2021). Hence, in order to further validate predictive models, an external validation is necessary using data from diverse US devices, clinical centers, and countries. Another limitation is the lack of histologically confirmed diagnoses among the conservatively treated patients. This implies that the model validation and comparison results presented above must be interpreted cautiously since we do not have access to the true disease status for all subjects. The image preprocessing pipeline could be improved further: currently, we discard scale information in the US images, making it impossible to detect the appendix diameter, a relevant sonographic sign of appendicitis (Reddan et al., 2016). Lastly, concepts could be modeled in a more fine-grained manner to incorporate physicians' uncertainty. Instead of just differentiating between the lack or presence of a finding, intermediate concept categories could be included by, for example, collecting data from multiple raters and considering discrepancies among them.

From the methodological perspective, we currently have a limited theoretical understanding of the (SS)MVCBMs. In particular, it would be desirable to explore the representations learned by SSMVCBMs and the identifiability of the ground-truth generative factors. Moreover, in the current implementation, it is not trivial to interpret the representations; thus, additional regularization may be necessary, such as rendering these representations disentangled.

Another potential improvement would be adopting a probabilistic approach to the concept and target variable prediction, facilitating more principled uncertainty estimation. As evidenced by the experiments, our predictive models could benefit from calibration. Explicit uncertainty modeling would allow for better-calibrated and more interpretable probabilistic predictions that could be utilized downstream to perform selective classification (Geifman and El-Yaniv, 2017) and uncertainty-based concept interventions (Shin et al., 2023). In practice, uncertainty in concept predictions could be modeled by adapting the proposed architecture with the modules from the stochastic segmentation networks (Larochelle et al., 2020) or probabilistic concept bottlenecks (Kim et al., 2023).

5. Conclusion and outlook

Motivated by the demand for model interpretability in biomedical applications, we investigated the use of concept bottleneck models for predicting the diagnosis, management and severity among pediatric patients with suspected appendicitis, leveraging abdominal ultrasound images. The densely annotated dataset used to develop the predictive models was made publicly available, and one of the models was deployed as a freely available demo web tool (https://papt.inf.ethz.ch/mvcbm). Methodologically, we introduced several enhancements to the conventional concept-based classification approach. Our proposed models can handle multiple views of the object of interest and insufficient concept sets. Overall, our experimental results suggest that the proposed methods can deliver competitive performance, while offering an alternative to black-box deep learning models and allowing for real-time interaction with the end user.

In future work, we aim to address several limitations outlined above. We plan to validate the predictive models externally on the data from a hospital located in another country. Various model design alterations, such as other choices of learnable fusion, further regularization of the learned representations, and uncertainty quantification, are also to be considered. Moreover, we recognize the significance of extending our investigation beyond the retrospective study. For instance, it would be interesting to explore the use of active learning to decide on the acquisition of US images and concept labels for each subject. From the clinical perspective, developed models should be extended to incorporate clinical and laboratory parameters and consider other conditions, such as COVID-19, during appendicitis. Additionally, we anticipate that using more refined definitions of the target variables could provide more insightful results, e.g. differentiating between subacute and acute appendicitis for the diagnosis and predicting the risk of secondary appendectomy for the management. Adjustments in the model architecture and the acquisition of a larger training dataset will facilitate the incorporation of the color Doppler images in the analysis, potentially making the prediction of the disease severity progression more accurate.

Declaration of competing interest

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

Data availability

The anonymized data are available on Zenodo at https://doi.org/10.5281/zenodo.7711412, and the code can be found in a GitHub repository at https://github.com/i6092467/semi-supervised-multiview-cbm.

Acknowledgments

We thank the members of the Medical Data Science group at ETH Zurich for stimulating discussions and feedback. Our gratitude goes to Dr. Johanna Joe and Dr. Markus Ebert from the Ultrasonography Center, KUNO, St. Hedwig Clinic Regensburg, for their help with the data acquisition. We are also sincerely thankful to Marcel Buehring and the IT Service Group of the Department of Computer Science, ETH Zurich, for their assistance with the deployment of the online prediction tool. RM was supported by the SNSF grant #320038189096, EO was supported by the SNSF grant P500PT-206746.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.103042.

References

Acharya, A., Markar, S.R., Ni, M., Hanna, G.B., 2016. Biomarkers of acute appendicitis: systematic review and cost–benefit trade-off analysis. Surg. Endosc. 31 (3), 1022–1031. http://dx.doi.org/10.1007/s00464-016-5109-1.

Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M., 2021. Representation learning with statistical independence to mitigate bias. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa, HI, USA, http://dx.doi.org/10.1109/wacv48630.2021.00256.

Akmese, O.F., Dogan, G., Kor, H., Erbay, H., Demir, E., 2020. The use of machine learning approaches for the diagnosis of acute appendicitis. Emerg. Med. Int. 2020, 1–8. http://dx.doi.org/10.1155/2020/7306435.

Alvarado, A., 1986. A practical score for the early diagnosis of acute appendicitis. Ann. Emerg. Med. 15 (5), 557–564. http://dx.doi.org/10.1016/S0196-0644(86)80993-3.

Andersson, R.E., 2006. The natural history and traditional management of appendicitis revisited: Spontaneous resolution and predominance of prehospital perforations imply that a correct diagnosis is more important than an early diagnosis. World J. Surg. 31 (1), 86–92. http://dx.doi.org/10.1007/s00268-006-0056-y.

Aydin, E., Türkmen, İ.U., Namli, G., Öztürk, Ç., Esen, A.B., Eray, Y.N., Eroğlu, E., Akova, F., 2020. A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children. Pediatr. Surg. Int. 36 (6), 735–742. http://dx.doi.org/10.1007/s00383-020-04655-7.

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443. http://dx.doi.org/10.1109/TPAMI.2018.2798607.

Bhangu, A., reide, K.S., Saverio, S.D., Assarsson, J.H., Drake, F.T., 2015. Acute appendicitis: modern understanding of pathogenesis, diagnosis, and management. Lancet 386 (10000), 1278–1287. http://dx.doi.org/10.1016/s0140-6736(15)00275-5.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/a:1010933404324.

Cheplygina, V., 2019. Cats or CAT scans: Transfer learning from natural or medical image source data sets? Curr. Opin. Biomed. Eng. 9, 21–27. http://dx.doi.org/10.1016/j.cobme.2018.12.005.

CODA Collaborative, 2020. A randomized trial comparing antibiotics with appendectomy for appendicitis. N. Engl. J. Med. 383 (20), 1907–1919. http://dx.doi.org/10.1056/nejmoa2014320.

Cogswell, M., Ahmed, F., Girshick, R.B., Zitnick, L., Batra, D., 2016. Reducing overfitting in deep networks by decorrelating representations. In: Bengio, Y., LeCun, Y. (Eds.), 4th International Conference on Learning Representations, ICLR 2016. http://dx.doi.org/10.48550/arXiv.1511.06068.

Deleger, L., Brodzinski, H., Zhai, H., Li, Q., Lingren, T., Kirkendall, E.S., Alessandrini, E., Solti, I., 2013. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. J. Am. Med. Inform. Assoc. 20 (e2), e212–e220. http://dx.doi.org/10.1136/amiajnl-2013-001962.

Dingemann, J., Ure, B., 2012. Imaging and the use of scores for the diagnosis of appendicitis in children. Eur. J. Pediatr. Surg. 22 (03), 195–200. http://dx.doi.org/10.1055/s-0032-1320017.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. http://dx.doi.org/10.48550/arXiv.1702.08608, arXiv:1702.08608.

Geifman, Y., El-Yaniv, R., 2017. Selective classification for deep neural networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 30, Curran Associates, Inc., pp. 4885—4894, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Geiger, D., Verma, T., Pearl, J., 1990. D-separation: From theorems to algorithms. In: Henrion, M., Shachter, R.D., Kanal, L.N., Lemmer, J.F. (Eds.), Uncertainty in Artificial Intelligence. In: Machine Intelligence and Pattern Recognition, vol. 10, North-Holland, pp. 139–148. http://dx.doi.org/10.1016/B978-0-444-88738-2.50018-X.

Gendel, I., Gutermacher, M., Buklan, G., Lazar, L., Kidron, D., Paran, H., Erez, I., 2011. Relative value of clinical, laboratory and imaging tools in diagnosing pediatric acute appendicitis. Eur. J. Pediatr. Surg. 21 (4), 229–233. http://dx.doi.org/10.1055/s-0031-1273702.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144. http://dx.doi.org/10.1145/3422622.

Gorter, R.R., Eker, H.H., Gorter-Stam, M.A.W., Abis, G.S.A., Acharya, A., Ankersmit, M., Antoniou, S.A., Arolfo, S., Babic, B., Boni, L., Bruntink, M., van Dam, D.A., Defoort, B., Deijen, C.L., DeLacy, F.B., Go, P.M., Harmsen, A.M.K., van den Helder, R.S., Iordache, F., Ket, J.C.F., Muysoms, F.E., Ozmen, M.M., Papoulas, M., Rhodes, M., Straatman, J., Tenhagen, M., Turrado, V., Vereczkei, A., Vilallonga, R., Deelder, J.D., Bonjer, J., 2016. Diagnosis and management of acute appendicitis. EAES consensus development conference 2015. Surg. Endosc. 30 (11), 4668–4690. http://dx.doi.org/10.1007/s00464-016-5245-7.

van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107. http://dx.doi.org/10.1158/0008-5472.can-17-0339.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. HeMIS: Hetero-modal image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, Athens, Greece, pp. 469–477. http://dx.doi.org/10.1007/978-3-319-46723-8_54.

Havasi, M., Parbhoo, S., Doshi-Velez, F., 2022. Addressing leakage in concept bottleneck models. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (Eds.), Advances in Neural Information Processing Systems. URL https://openreview.net/forum?id=tglniD_fn9.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, http://dx.doi.org/10.1109/cvpr.2016.90.

Hsieh, C.-H., Lu, R.-H., Lee, N.-H., Chiu, W.-T., Hsu, M.-H., Li, Y.-C.J., 2011. Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery 149 (1), 87–93. http://dx.doi.org/10.1016/j.surg.2010.03.023.

Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A., 2020. Variational autoencoders and nonlinear ICA: A unifying framework. In: Chiappa, S., Calandra, R. (Eds.), Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol. 108, PMLR, Virtual, pp. 2207–2217.

Kim, E., Jung, D., Park, S., Kim, S., Yoon, S., 2023. Probabilistic concept bottleneck models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.), Proceedings of the 40th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 202, PMLR, pp. 16521–16540, URL https://proceedings.mlr.press/v202/kim23g.html.

Kiss, N., Minderjahn, M., Reismann, J., Svensson, J., Wester, T., Hauptmann, K., Schad, M., Kallarackal, J., von Bernuth, H., Reismann, M., 2021. Use of gene expression profiling to identify candidate genes for pretherapeutic patient classification in acute appendicitis. BJS Open 5 (1), http://dx.doi.org/10.1093/bjsopen/zraa045.

Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P., 2020. Concept bottleneck models. In: Daumé III, H., Singh, A. (Eds.), Proceedings of the 37th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 119, PMLR, Virtual, pp. 5338–5348.

Kryzauskas, M., Danys, D., Poskus, T., Mikalauskas, S., Poskus, E., Jotautas, V., Beisa, V., Strupas, K., 2016. Is acute appendicitis still misdiagnosed? Open Med. 11 (1), 231–236. http://dx.doi.org/10.1515/med-2016-0045.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems. vol. 34, Curran Associates, Inc., Red Hook, NY, United States, pp. 16451–16467.

Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2009. Attribute and simile classifiers for face verification. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, Kyoto, Japan, pp. 365–372. http://dx.doi.org/10.1109/ICCV.2009.5459250.

Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, USA, http://dx.doi.org/10.1109/CVPR.2009.5206594.

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B., 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems. vol. 33, Curran Associates, Inc., pp. 12756–12767, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/95f8d9901ca8878e291552f001f67692-Paper.pdf.

Ma, C., Guo, Y., Yang, J., An, W., 2019. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. IEEE Trans. Multimed. 21 (5), 1169–1182. http://dx.doi.org/10.1109/TMM.2018.2875512.

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., Pan, W., 2021. Promises and pitfalls of black-box concept learning models. http://dx.doi.org/10.48550/arXiv.2106.13314, arXiv:2106.13314.

Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., Vogt, J.E., 2021. Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. Front. Pediatr. 9, http://dx.doi.org/10.3389/fped.2021.662183.

Marconato, E., Passerini, A., Teso, S., 2022. GlanceNets: Interpretable, leak-proof concept-based models. http://dx.doi.org/10.48550/arXiv.2205.15612, arXiv:2205.15612.

Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A., 2021. Do concept bottleneck models learn as intended?. http://dx.doi.org/10.48550/arXiv.2105.04289, arXiv:2105.04289.

Mostbeck, G., Adam, E.J., Nielsen, M.B., Claudon, M., Clevert, D., Nicolau, C., Nyhsen, C., Owens, C.M., 2016. How to diagnose acute appendicitis: ultrasound first. Insights Imaging 7 (2), 255–263. http://dx.doi.org/10.1007/s13244-016-0469-6.

Ohba, G., Hirobe, S., Komori, K., 2016. The usefulness of combined b mode and Doppler ultrasonography to guide treatment of appendicitis. Eur. J. Pediatr. Surg. 26 (06), 533–536. http://dx.doi.org/10.1055/s-0035-1570756.

Park, N.H., Oh, H.E., Park, H.J., Park, J.Y., 2011. Ultrasonography of normal and abnormal appendix in children. World J. Radiol. 3 (4), 85–91. http://dx.doi.org/10.4329/wjr.v3.i4.85.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., dÁlché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 32, Curran Associates, Inc., Red Hook, NY, United States.

Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., Sun, Q., Lu, L., Shung, K.K., 2021. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat. Biomed. Eng. 5 (6), 522–532. http://dx.doi.org/10.1038/s41551-021-00711-2.

Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C.P., Lungren, M.P., Ng, A.Y., Patel, B.N., 2020. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. Sci. Rep. 10 (1), http://dx.doi.org/10.1038/s41598-020-61055-6.

Reddan, T., Corness, J., Mengersen, K., Harden, F., 2016. Ultrasound of paediatric appendicitis and its secondary sonographic signs: providing a more meaningful finding. J. Med. Radiat. Sci. 63 (1), 59–66. http://dx.doi.org/10.1002/jmrs.154.

Reismann, J., Romualdi, A., Kiss, N., Minderjahn, M.I., Kallarackal, J., Schad, M., Reismann, M., 2019. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. PLoS One 14 (9), e0222030. http://dx.doi.org/10.1371/journal.pone.0222030.

Roig Aparicio, P., Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., Vogt, J.E., 2021. Learning medical risk scores for pediatric appendicitis. In: 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, Pasadena, CA, USA, http://dx.doi.org/10.1109/ICMLA52953.2021.00243.

RSGobotWMR, C., 2020. Appendicitis risk prediction models in children presenting with right iliac fossa pain (RIFT study): a prospective, multicentre validation study. Lancet Child Adolesc. Health 4 (4), 271–280. http://dx.doi.org/10.1016/S2352-4642(20)30006-7.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215. http://dx.doi.org/10.1038/s42256-019-0048-x.

Samuel, M., 2002. Pediatric appendicitis score. J. Pediatr. Surg. 37 (6), 877–881. http://dx.doi.org/10.1053/jpsu.2002.32893.

Saverio, S.D., Birindelli, A., Kelly, M.D., Catena, F., Weber, D.G., Sartelli, M., Sugrue, M., Moya, M.D., Gomes, C.A., Bhangu, A., Agresta, F., Moore, E.E., Soreide, K., Griffiths, E., Castro, S.D., Kashuk, J., Kluger, Y., Leppaniemi, A., Ansaloni, L., Andersson, M., Coccolini, F., Coimbra, R., Gurusamy, K.S., Campanile, F.C., Biffl, W., Chiara, O., Moore, F., Peitzman, A.B., Fraga, G.P., Costa, D., Maier, R.V., Rizoli, S., Balogh, Z.J., Bendinelli, C., Cirocchi, R., Tonini, V., Piccinini, A., Tugnoli, G., Jovine, E., Persiani, R., Biondi, A., Scalea, T., Stahel, P., Ivatury, R., Velmahos, G., Andersson, R., 2016. WSES Jerusalem guidelines for diagnosis and treatment of acute appendicitis. World J. Emerg. Surg. 11 (1), http://dx.doi.org/10.1186/s13017-016-0090-5.

Sawada, Y., Nakamura, K., 2022. Concept bottleneck model with additional unsupervised concepts. IEEE Access 10, 41758–41765. http://dx.doi.org/10.1109/ACCESS.2022.3167702.

Shin, S., Jo, Y., Ahn, S., Lee, N., 2023. A closer look at the intervention procedure of concept bottleneck models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.), Proceedings of the 40th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 202, PMLR, pp. 31504–31520, URL https://proceedings.mlr.press/v202/shin23a.html.

Stiel, C., Elrod, J., Klinke, M., Herrmann, J., Junge, C.-M., Ghadban, T., Reinshagen, K., Boettcher, M., 2020. The modified Heidelberg and the AI appendicitis score are superior to current scores in predicting appendicitis in children: A two-center cohort study. Front. Pediatr. 8, http://dx.doi.org/10.3389/fped.2020.592892.

Suzuki, M., Matsuo, Y., 2022. A survey of multimodal deep generative models. Adv. Robot. 36 (5–6), 261–278. http://dx.doi.org/10.1080/01691864.2022.2035253.

Svensson, J., Hall, N., Eaton, S., Pierro, A., Wester, T., 2012. A review of conservative treatment of acute appendicitis. Eur. J. Pediatr. Surg. 22 (03), 185–194. http://dx.doi.org/10.1055/s-0032-1320014.

Svensson, J.F., Patkova, B., Almström, M., Naji, H., Hall, N.J., Eaton, S., Pierro, A., Wester, T., 2015. Nonoperative treatment with antibiotics versus surgery for acute nonperforated appendicitis in children. Ann. Surg. 261 (1), 67–71. http://dx.doi.org/10.1097/sla.0000000000000835.

Taeb, A., Ruggeri, N., Schnuck, C., Yang, F., 2022. Provable concept learning for interpretable predictions using variational autoencoders. http://dx.doi.org/10.48550/arXiv.2204.00492, arXiv:2204.00492.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P., 2020. What makes for good views for contrastive learning? In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems. vol. 33, Curran Associates, Inc., Red Hook, NY, United States, pp. 6827–6839.

Wang, Y., Choi, E.J., Choi, Y., Zhang, H., Jin, G.Y., Ko, S.-B., 2020. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. Ultras. Med. Biol. 46 (5), 1119–1132. http://dx.doi.org/10.1016/j.ultrasmedbio.2020.01.001.

Wier, L.M., Yu, H., Owens, P.L., Washington, R., 2013. Overview of Children in the Emergency Department, 2010: Statistical Brief #157. Agency for Healthcare Research and Quality, Rockville, MD, USA.

Xia, J., Wang, Z., Yang, D., Li, R., Liang, G., Chen, H., Heidari, A.A., Turabieh, H., Mafarja, M., Pan, Z., 2022. Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. Comput. Biol. Med. 143, 105206. http://dx.doi.org/10.1016/j.compbiomed.2021.105206.

Xian, Y., Lampert, C.H., Schiele, B., Akata, Z., 2019. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. 41 (9), 2251–2265. http://dx.doi.org/10.1109/tpami.2018.2857768.

Xu, C., Tao, D., Xu, C., 2013. A survey on multi-view learning. http://dx.doi.org/10.48550/arXiv.1304.5634, arXiv:1304.5634.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., Ravikumar, P., 2020. On completeness-aware concept-based explanations in deep neural networks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems. vol. 33, Curran Associates, Inc., Vancouver, Canada, pp. 20554–20565.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, http://dx.doi.org/10.1109/cvpr.2018.00577.

Yuksekgonul, M., Wang, M., Zou, J., 2022. Post-hoc concept bottleneck models. http://dx.doi.org/10.48550/arXiv.2205.15480, arXiv:2205.15480.