Multiview Concept Bottleneck Models Applied to Diagnosing Pediatric Appendicitis

Ugne Klimiene [*1] **Ričards Marcinkevičs** [*1] **Patricia Reis Wolfertstetter** [2] **Ece Ozkan** [1] **Alyssia Paschke** [3] **David Niederberger** [1] **Sven Wellmann** [2] **Christian Knorr** [2] **Julia E. Vogt** [1]

Abstract

Arguably, interpretability is one of the guiding principles behind the development of machine-learning-based healthcare decision support tools and computer-aided diagnosis systems. There has been a renewed interest in interpretable classification based on high-level concepts, including, among other model classes, the re-exploration of concept bottleneck models. By their nature, medical diagnosis, patient management, and monitoring require the assessment of multiple views and modalities to form a holistic representation of the patient's state. For instance, in ultrasound imaging, a region of interest might be registered from multiple views that are informative about different sets of clinically relevant features. Motivated by this, we extend the classical concept bottleneck model to the multiview classification setting by representation fusion across the views. We apply our multiview concept bottleneck model to the dataset of ultrasound images acquired from a cohort of pediatric patients with suspected appendicitis to predict the disease. The results suggest that auxiliary supervision from the concepts and aggregation across multiple views help develop more accurate and interpretable classifiers.

1. Introduction

One of the conventional models for interpretable classification (Doshi-Velez & Kim, 2017; Rudin, 2019) is concept bottleneck (Kumar et al., 2009; Lampert et al., 2009; Koh et al., 2020): (*i*) based on features $\boldsymbol{x}$, a vector of human-understandable concepts $\boldsymbol{c}$ is predicted; (*ii*) concepts $\boldsymbol{c}$ are then used to predict the label y. More formally, the prediction made by a concept bottleneck model is given by

$$\hat{y} = f_{\boldsymbol{\theta}}\left(\boldsymbol{g}_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right), \tag{1}$$

where $\boldsymbol{g}_{\boldsymbol{\phi}}(\cdot)$ is a neural network parameterised by $\boldsymbol{\phi}$, mapping inputs $\boldsymbol{x}$ to the predicted concepts $\hat{\boldsymbol{c}}$, and $f_{\boldsymbol{\theta}}(\cdot)$ maps $\hat{\boldsymbol{c}}$ to the predicted label $\hat{y}$. For the sake of convenience, we will refer to $\boldsymbol{g}_{\boldsymbol{\phi}}(\cdot)$ as the concept model and to $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ as the target model, similar to Lockhart et al. (2022). When the model given by Equation 1 is deployed, a human user can interpret and interact with the model's predictions by inspecting and editing the concepts. In this work, we extend the concept bottleneck models (CBM), as described by Koh et al. (2020), to the *multiview* classification setting pertinent to computer-aided diagnosis based on medical imaging data.

In the multiview setting, instead of observing a single set of features $\boldsymbol{x}_i$ for each data point $1 \leq i \leq N$, we are given a sequence of V views $\{\boldsymbol{x}_i^v\}_{v=1}^V$ (Nie et al., 2018). Let us consider a simple example depicted in Figure 1, wherein we are given images corresponding to $V = 3$ views of a single bird. Assuming the task is to predict bird species based on the images and a set of attributes, similar to the Caltech-UCSD Birds (Welinder et al., 2010), we must remark that not every concept c_j may be identifiable from every view. For instance, in Figure 1, view I is not informative about the beak shape, whereas the back pattern cannot be detected from view III. We will refer to this phenomenon as "*partial observability*". In such a multiview setting, representations need to be aggregated across the views to predict the full set of concepts. Moreover, the dimensionality of $\boldsymbol{x}_i^v$ might vary across $1 \leq v \leq V$, and some views might be missing; therefore, trivially concatenating features across the views may not be a satisfactory solution. The multiview concept bottleneck model presented in this paper allows handling variability and missingness in a principled manner.

Patient screening and diagnosis based on medical imaging data often give rise to the multiview setting outlined above. For example, the risk of breast cancer may be assessed based on multiview and multimodal ultrasound (US) images of lesions (Wang et al., 2020a; Qian et al., 2021), including transversal and longitudinal views of B-mode,

[*]Equal contribution [1]ETH Zurich, Switzerland [2]University Children's Hospital Regensburg, Germany [3]University of Regensburg, Germany. Correspondence to: Ričards Marcinkevičs <ricards.marcinkevics@inf.ethz.ch>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).

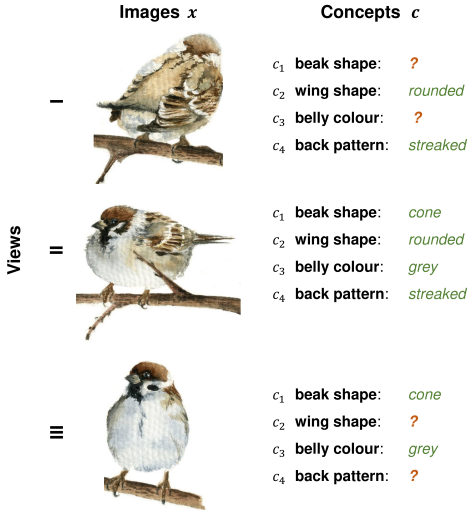

Figure 1. Concept-based classification in the multiview setting. Three views (I, II, III) of a bird are observed (*left*): not *all* concepts (*right*) are identifiable from *every* view.

colour Doppler, and elastography images. In this work, we leverage the proposed model for *interpretable* prediction of the diagnosis among pediatric patients with suspected appendicitis. Appendicitis is one of the commonest causes of abdominal pain and one of the most frequent diagnoses resulting in hospital admissions of patients under 18 (Wier et al., 2013). In particular, we investigate the use of the multiview abdominal US data to predict this disease. Even though the US has lower sensitivity and specificity than computed tomography and magnetic resonance imaging, due to the absence of ionising radiation and it's real-time technology, US has been advocated to be the preferred imaging modality for diagnosing acute appendicitis (Mostbeck et al., 2016). To the best of our knowledge, this is one of the first studies focusing on the machine-learning-based prediction of pediatric appendicitis directly utilising US images.

Our Contributions The contributions of this work are threefold: (*i*) a novel extension of concept bottleneck models to multiview learning under partial observability of the views and the concepts from individual views; (*ii*) application of the proposed model to a multiview abdominal US dataset from a tertiary care hospital to predict the diagnosis in pediatric patients with suspected appendicitis; (*iii*) empirical comparison to several statistical and black-box ML approaches.

2. Related Work

Concept-based Models Many recent works have re-explored concept-based prediction (Koh et al., 2020; Chen et al., 2020; Marcos et al., 2021; Losch et al., 2021), i.e. prediction based on high-level semantic features that are usually provided as auxiliary supervision at the training time. Concept-based models are deemed interpretable since concepts can be inspected alongside the model's outputs and viewed as "explanations". In addition, Koh et al. (2020) make their concept bottleneck models *intervenable*, i.e., at the test time, the concepts may be edited by a human expert to change the model's predictions. Several further efforts have been made to better understand and address the limitations of concept bottlenecks (Mahinpei et al., 2021; Margeloiu et al., 2021; Lockhart et al., 2022; Sawada & Nakamura, 2022), focusing on mitigating information leakage, improving concept intervenability, and extending the model to semi-supervised representation learning. Another promising related line of work has focused on testing for associations and extracting concepts from already trained networks *post hoc* (Kim et al., 2018; Yeh et al., 2020). As opposed to CBMs, the latter category of methods only allows explaining predictions but not intervening on them.

Multiview Learning Multiview learning (Xu et al., 2013) is tailored to the data comprising multiple views, essentially, feature subsets, of the same source object (see Figure 1). Multiview data naturally arise from various healthcare settings, including but not limited to ultrasound imaging (Wang et al., 2020a; Qian et al., 2021), multiomics analysis (Nguyen & Wang, 2020), and neuroimaging (Zhang et al., 2018). Recently, connections have been drawn between multiview and contrastive and self-supervised learning (Tian et al., 2020; Tsai et al., 2021). Another related field is multimodal learning (Baltrusaitis et al., 2019) which develops models combining multiple heterogeneous modalities, e.g. images and text.

Machine Learning for Appendicitis There is extensive research on leveraging machine learning models to diagnose and manage patients with suspected appendicitis. With few exceptions, most models either utilise simple clinical and laboratory data, rely on hand-crafted US annotations, or require more expensive imaging modalities, such as computed tomography (CT). Many predictive models have been designed specifically for pediatric patients (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), relying only on tabular data. By contrast, Deleger et al. (2013) leveraged natural language processing to analyse electronic health record contents and assess the risk of acute appendicitis in ED patients. For adult population, Hsieh et al. (2011) have applied logistic regression, SVMs, random forests, and neural networks to demographic, clinical, and laboratory variables. Utilising pretrained models, Rajpurkar et al. (2020) developed a 3D CNN for classifying patients on a small dataset of CT exams. To the best of our knowledge, direct processing of abdominal US images remains an under-explored topic.

3. Method

Throughout this paper, we will assume the following setting and notation. Consider being given 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 $1 \leq i \leq N$. In this work, we concentrate on the scenario wherein all views are given by images that can be preprocessed and rescaled into the same dimensionality. Nevertheless, our approach can be readily extended to multiple heterogeneous data types.

Without loss of generality, we focus on the data exhibiting characteristics described informally below. (*i*) *Partial observability*: not all concepts are identifiable from all views (Figure 1). (*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 (Section 5) and support some of the modelling choices described below.

3.1. Multiview Concept Bottleneck Model

We now introduce an extension of the concept bottleneck models (Koh et al., 2020) to the multiview setting outlined above. Henceforth, we refer to this extension as *multiview concept bottleneck model* (MVCBM). Figure 2 provides a schematic summary of the MVCBM architecture discussed in detail below. Equation 2 contains equations for a forward pass of the model. In brief, MVCBM consists of the following modules: (*i*) per-view feature extraction; (*ii*) feature fusion; (*iii*) concept prediction; and (*iv*) label prediction.

Step *i*: Feature Extraction Given an image sequence $\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}$ representing ordered views, we first encode each image into a lower-dimensional view-specific representation. We use a *shared* encoder neural network, denoted by $\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$, across all views. Weight sharing is justified by the view homogeneity assumption and could be helpful in smaller datasets with high missingness of views. On the other hand, if the dataset is relatively regular and the differences across views are considerable, especially in the multimodal setting, one could train a dedicated encoder for each view. In practice, it may be prudent to use a pretrained model to initialise $\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$, e.g. the use of ResNet (He et al., 2016) and VGG (Simonyan & Zisserman, 2015) architectures pretrained on natural images is common in medical imaging applications (Cheplygina, 2019). As a result of this step, we obtain a sequence of view-specific features given by $\boldsymbol{h}_i^v = \boldsymbol{h}_{\boldsymbol{\psi}}\left(\boldsymbol{x}_i^v\right)$, for $1 \leq v \leq V_i$ and $1 \leq i \leq N$ (Equation 2a).

Step *ii*: Feature Fusion To handle multiple views, we need to fuse, i.e. aggregate, them within the model. MVCBM follows a *hybrid* fusion approach (Baltrusaitis et al., 2019): rather than concatenating views at the input level or training an ensemble of view-specific models; we aggregate intermediate view-specific features $\boldsymbol{h}_i^v$ from the previous step within a single neural network (Equation 2b). Although there are many viable fusion functions, in our context, the fusion must handle varying numbers of views per data point. As a naïve approach, we consider taking an 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, $\bar{\boldsymbol{h}}_i = \boldsymbol{r}_{\boldsymbol{\xi}}\left(\{\boldsymbol{h}_i^v\}_{v=1}^{V_i}\right)$, where $\bar{\boldsymbol{h}}_i$ 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 utilise this trick in multiview 3D shape recognition, we combine view-specific representations via a long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997). In particular, we set the aggregated representation $\bar{\boldsymbol{h}}_i$ to the last hidden state of the view sequence, i.e. at 'time 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, which we leave for the future work.

Steps *iii–iv*: Concept and Label Prediction The last two steps 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 $\boldsymbol{s}_{\boldsymbol{\zeta}}(\cdot)$ parameterised by $\boldsymbol{\zeta}$ (Equation 2c). 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 $\hat{y}$ (Equation 2d). The output activation should be chosen based on the downstream task: classification or regression. In this work, we focus exclusively on classification.

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

(*i*) Feature extraction:

$$\boldsymbol{h}_i^v = \boldsymbol{h}_{\boldsymbol{\psi}}\left(\boldsymbol{x}_i^v\right), \; 1 \leq v \leq V_i, \tag{2a}$$

(*ii*) Feature fusion:

$$\bar{\boldsymbol{h}}_i = \boldsymbol{r}_{\boldsymbol{\xi}}\left(\{\boldsymbol{h}_i^v\}_{v=1}^{V_i}\right), \tag{2b}$$

(*iii*) Concept prediction:

$$\hat{\boldsymbol{c}}_i = \boldsymbol{s}_{\boldsymbol{\zeta}}\left(\bar{\boldsymbol{h}}_i\right), \tag{2c}$$

(*iv*) Label prediction:

$$\hat{y}_i = f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{c}}_i\right). \tag{2d}$$

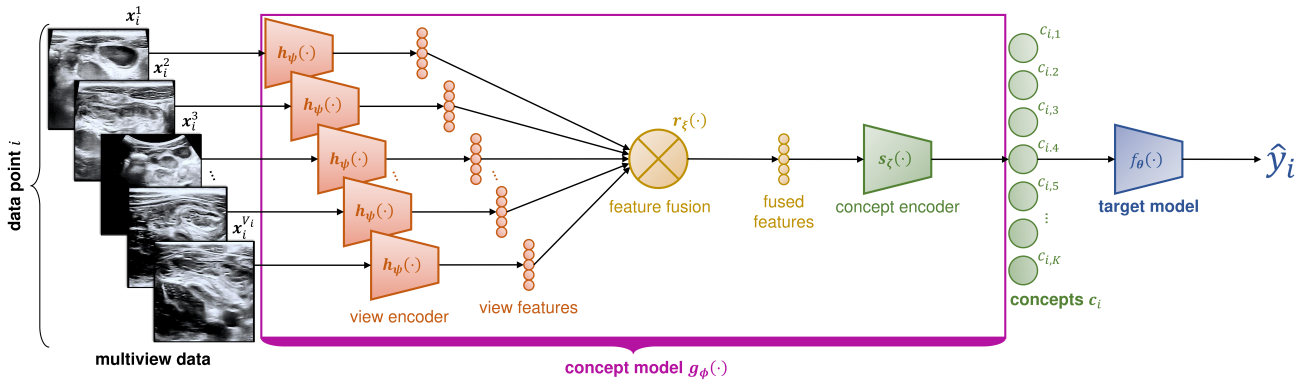

Figure 2. Schematic summary of the proposed multiview concept bottleneck model: ***(i)*** multiview data are mapped to features using a *shared* encoder; ***(ii)*** features are fused across the views; ***(iii)*** concepts $\boldsymbol{c}$ are predicted based on the fused features, ***(iv)*** final prediction is made, using the target model $f_{\boldsymbol{\theta}}(\cdot)$. Observe that steps (*i*)-(*iii*) comprise the concept model $\boldsymbol{g}_{\boldsymbol{\phi}}(\cdot)$.

Observe that, in MVCBM, the concept model $\boldsymbol{g}_{\boldsymbol{\phi}}(\cdot)$ is composed of several steps, and its parameters $\boldsymbol{\phi}$ correspond to $\{\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\zeta}\}$ (cf. Equations 1 and 2).

Loss Function and Optimisation Koh et al. (2020) discuss independent, sequential, and joint optimisation procedures for the CBMs. In this work, we focus on the sequential and joint approaches since, according to the original paper, they offer a more balanced trade-off between predictive performance and intervenability.

In particular, in the *sequential* training, we first optimise the concept model parameters:

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^t w_i^{c_k} \mathcal{L}^{c_k}(\hat{c}_{i,k}, c_{i,k}), \tag{3}$$

where $\mathcal{L}^{c_k}(\cdot, \cdot)$ is the loss function for the k-th 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 k-th concept for the i-th 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 k-th concept and w_i^t for the target variable of the i-th 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 normalised inverse frequencies of the variable classes. In the next step, parameters $\hat{\boldsymbol{\phi}}$ are frozen, and the parameters of the target model $f_{\boldsymbol{\theta}}$ are optimised:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} w_i^t \mathcal{L}^t\left(f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{c}}_i\right), y_i\right), \tag{4}$$

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 Equations 3 and 4 into a single objective:

$$\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} w_i^t \mathcal{L}^t(\hat{y}_i, y_i) + \alpha \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^t w_i^{c_k} \mathcal{L}^{c_k}(\hat{c}_{i,k}, c_{i,k}) \right\}, \tag{5}$$

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

3.2. Implementation Details

We implemented MVCBM in PyTorch (v 1.10.2) (Paszke et al., 2017). Across all experiments, we fine-tuned pre-trained ResNet-18 (He et al., 2016) as the shared view encoder $\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$; the arithmetic mean and LSTM were used as the fusion function $\boldsymbol{r}_{\boldsymbol{\xi}}(\cdot)$; for the concept encoder $\boldsymbol{s}_{\boldsymbol{\zeta}}(\cdot)$ and target model $f_{\boldsymbol{\theta}}(\cdot)$, we utilised multilayer perceptrons with ReLU hidden activations. Detailed architecture specification and training procedure are provided in Appendix A.

4. Cohort, Data and Evaluation

The purpose of our experiments was twofold: (*i*) provide a proof of concept for the introduced multiview extension of the CBMs on natural images and (*ii*) apply MVCBM to a real-world dataset of abdominal US images acquired from patients with suspected appendicitis. This section contains a brief overview of the datasets, baselines, and evaluation procedures employed in the experiments (Section 5).

4.1. Datasets

Multiview Animals with Attributes (MVAwA) To showcase the utility of the multiview approach to the concept-based classification and test the feasibility of our MVCBM

model, we adapted a popular attribute-based classification dataset *Animals with Attributes 2* (AwA) (Xian et al., 2019; Lampert et al., 2009) to the multiview setting. The original AwA consists of 37,322 images of 50 animal classes with 85 binary-valued concepts, i.e. attributes. Similar to the UCSD Birds experiment by Koh et al. (2020), the concepts are labelled per class and *not* per data point. We extended AwA by randomly cropping $V_i = 4$ patches, 60×60 px^2 big, from each image i to produce multiple "views" (Figure 5, Appendix B). Note that while the concepts are only partially observable from individual images, there is no ordering among the views in this dataset, and, for simplicity, we enforce the same V_i for each data point. Nevertheless, compared with the original AwA, the classification problem becomes remarkably more challenging (Section 5.1).

Appendicitis The appendicitis dataset includes 275 patients aged from 0 to 18 years admitted with abdominal pain and suspected appendicitis to the surgery department of the tertiary care Children's Hospital St. Hedwig in Regensburg, Germany, 2016–2018. At our disposal, we had 42 demographic, clinical, scoring, laboratory, and ultrasound predictor variables and 824 ultrasound images. Each subject corresponded to a single data point with views loosely ordered based on the examination time. Diagnosis (*appendicitis* vs. *no appendicitis*) was used as the target. Its categories were mildly imbalanced: 62 vs. 38%. Note that this variable is a proxy for the true diagnosis since there was no cohort-wide histological confirmation. In patients treated conservatively, diagnosis was assigned by physicians based on an *ad hoc* criterion: an Alvarado or pediatric appendicitis score of $\geq$ 4 and an appendix diameter of $\geq$ 6 mm. We selected eleven binary predictors (see the list and the descriptive statistics in Table 6, Appendix C) as concepts for the MVCBM. The selection criteria were as follows: (*i*) the predictor variable had to be detectable from ultrasound images, as confirmed by a qualified physician, and (*ii*) the variable had to had been collected preoperatively. Missing concept values were imputed with negative outcomes since missingness usually indicated the absence of the finding.

4.2. Image Preprocessing

For the MVAwA, all views were rescaled between 0 and 1 and resized to 224×224 px^2. No augmentation was applied to the inputs during training. For the ultrasound images, we employed a generative inpainting model DeepFill (Yu et al., 2018) to mask and fill in the graphical user interface of the US device, markers, distance measurements, and other annotations in the original B-mode images. Subsequently, they were cropped to 400×400 px^2 dimensions using zero padding when needed. Finally, the contrast limited histogram equalisation (CLAHE) was performed, images were scaled to the range between 0 and 1, and normalised as required by the pretrained PyTorch models. Figure 3 shows an example of one image from the appendicitis dataset before and after preprocessing. The models trained on the appendicitis data used extensive on-the-fly augmentation with one randomly chosen transformation per image (Appendix A).

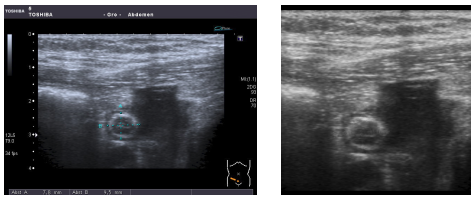

Figure 3. An example of one view from the appendicitis dataset: the original ultrasound image (*left*) contains graphical interface elements and expert-made markers, whereas the preprocessed image (*right*) has been inpainted, cropped, and padded.

4.3. Ablations, Baselines and Evaluation

We compared several variations of the proposed multiview concept bottleneck. Namely, we trained models using the sequential (*MVCBM-seq*) and joint (*MVCBM-joint*) optimisation procedures (Equations 3, 4, and 5). Moreover, two fusion functions were investigated: the arithmetic mean (*MVCBM-avg*) and LSTM (*MVCBM-LSTM*). Last but not least, similar to Koh et al. (2020), we experimented with intervening on the concept bottleneck by replacing the predicted concept values with the ground truth at test time (*intervened MVCBM*). The goal was to investigate whether a medical practitioner utilising our model could improve its predictions interactively. In particular, for a data point $1 \leq i \leq N$, the updated prediction after the intervention on the concepts from a subset $\mathcal{S} \subseteq \{1, ..., K\}$ is given by

$$\hat{y}_i^{\mathcal{S}} = f_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{c}}_{\{1,\dots,K\}\setminus\mathcal{S}}, \boldsymbol{c}_{\mathcal{S}}\right), \tag{6}$$

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 $f_{\hat{\boldsymbol{\theta}}}(\cdot)$.

For the appendicitis dataset, we benchmarked the performance of MVCBM against five baseline models: *Radiomics-RF*, *ResNet-18*, *MVBM-avg*, *MVBM-LSTM*, and *US-MLP* (Table 2). The first two baselines handle multiview data naïvely: they predict each view's label and return the average prediction across views for each data point. Concretely, the *Radiomics-RF* baseline extracts 100 radiomic features (van Griethuysen et al., 2017; Wang et al., 2020b) from every view. It then fits a random forest classifier to predict the target, whereas *ResNet-18* fine-tunes the ResNet-18 pretrained on the 1000-class ImageNet dataset. *MVBM-avg* and *MVBM-LSTM* correspond to *MVCBM-avg* and *MVCBM-LSTM*, respectively, and follow the architecture detailed in Figure 2, except that the bottleneck layer is unsupervised,

i.e. these models are not directly interpretable or intervenable. Finally, *US-MLP* refers to the multilayer perceptron, structurally identical to the one used by MVCBM as $f_{\boldsymbol{\theta}}(\cdot)$, predicting the target label from the *ground truth* concept values in contrast to the MVCBM that utilises predicted concept values. Informally, this baseline defines an upper bound on the performance of purely concept-based predictions assuming no information leakage through concept nodes (Lockhart et al., 2022). For the MVAwA, we, in addition, included the vanilla single-view CBM (*Single-CBM*) as a baseline to demonstrate the utility of multiview learning. For reference, we also report the expected performance of a random guess (*Random*), i.e. a *fair* coin flip. We used cross-validation (CV) for model comparison, evaluating the accuracy (ACC), balanced accuracy (BA), macro-averaged F1 score, and areas under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves for both concept and target predictions.

Model	Target		Concepts	
	ACC	BA	AUROC	AUPR
Random	0.02	0.02	0.50	0.36
Single-CBM-seq	0.22±0.07	0.18±0.06	0.86±0.00	0.75±0.00
-joint	0.13±0.07	0.12±0.05	0.86±0.00	0.76±0.00
MVBM-avg	0.35±0.13	0.30±0.11	—	—
-LSTM	0.28±0.11	0.23±0.09	—	—
MVCBM-seq-avg	*0.38±0.11*	*0.35±0.09*	**0.96±0.00**	**0.92±0.01**
-seq-LSTM	**0.41±0.09**	**0.36±0.08**	*0.95±0.00*	0.90±0.00
-joint-avg	0.25±0.08	0.23±0.08	**0.96±0.01**	**0.92±0.02**
-joint-LSTM	0.26±0.09	0.23±0.08	**0.96±0.00**	*0.91±0.00*

Table 1. Model performance comparison for the target and concept prediction on the MVAwA. Metrics are reported as averages and standard deviations across five folds of Monte Carlo cross-validation. AUROCs and AUPRs were averaged across concepts. **Bold** indicates the best result, *italics* indicates the second best.

5. Results

5.1. MVAwA

To test the feasibility of our model, we first applied it to the toy multiview adaptation of the AwA dataset. Note that these results are not comparable with the classical AwA since we crop the images into relatively small patches and do not consider a zero-shot learning scenario. Table 1 provides the results for different configurations of the MVCBM and several baselines. Observe that the multiview approaches outperform the vanilla CBM trained on a single view w.r.t. both predicting the target and concepts. It appears that supervision from the concepts does not hurt the performance since MVCBM models are on par with or even better than unsupervised bottlenecks (MVBM). As expected, we observe no considerable difference between average- and LSTM-based fusion, likely because, for MVAwA, the views are exchangeable. For predicting the target, jointly trained models perform, on average, worse than sequentially trained ones. We attribute this to the choice of the parameter α (Equation 5), which was kept at 1.0 across all experiments. Moreover, in practice, we observed that the joint optimisation requires careful hyperparameter tuning (Appendix A). Overall, the results agree with our expectations and suggest that MVCBM can effectively aggregate information across multiple views, improving the target and concept prediction.

5.2. Application to Pediatric Appendicitis

Predicting Diagnosis Table 2 shows the results for predicting the appendicitis diagnosis. All models are organised into three groups: *baselines*, *MVCBMs*, and *intervened MVCBMs*. Firstly, it is encouraging that all models perform systematically better than a random guess. Generally, hybrid fusion models (MVBM and MVCBM) perform better than the late-fusion-based model with a ResNet-18 backbone. Furthermore, observe that the models with the LSTM-based fusion consistently perform better than those using the simple arithmetic mean. We attribute this to the LSTM network leveraging temporal and spatial dependencies among the US images of a single subject. We have also compared sequential and joint optimisation procedures. Interestingly, jointly trained models, being more troublesome to optimise (Appendix A), exhibit results similar to those of the sequentially trained ones. Comparing MVCBMs with the corresponding MVBMs, we observe a trade-off between the model's predictive accuracy and interpretability, but only for the average-based fusion. Indeed, the performance of LSTM-based models does not suffer from the addition of a supervised concept bottleneck. Finally, the best configuration of the MVCBMs, namely, the sequentially trained LSTM-based model, outperforms all baselines except the US-MLP, which, in contrast to MVCBM, requires the costly acquisition of ultrasound variables by a medical specialist.

Predicting Concepts In addition, we investigated whether, next to their predictive performance, MVCBMs can capture the concepts accurately. Table 3 summarises the concept predictive performance w.r.t. AUROCs and AUPRs (see an extended version in Appendix D). For most concepts, sequentially trained MVCBMs achieve better performance overall. On the other hand, for a fixed optimisation procedure (sequential/joint) and the majority of concepts, LSTM-based image feature fusion is more effective than averaging. This is consistent with the results above: the seq-LSTM model is superior to the other MVCBMs (Table 2). Observe that concepts *thickening of the bowel wall* (c_3), *coprostasis* (c_7), and *gynaecological findings* (c_{11}) are especially hard to predict due to their highly imbalanced class distributions (Appendix C). On the contrary, *surrounding tissue reaction* (c_1), *visibility of the appendix* (c_5), and *pathological lymph nodes* (c_6) are predicted satisfactorily. They are known to be secondary signs of appendicitis (Reddan et al., 2016).

Group	Model	ACC	Macro F1	AUROC	AUPR
Baselines	Random	0.50	0.49	0.50	0.62
	Radiomics-RF	0.69±0.04	0.61±0.05	0.66±0.07	0.74±0.04
	ResNet-18	0.58±0.05	0.54±0.01	0.60±0.05	0.71±0.05
	MVBM-avg	0.65±0.05	0.63±0.07	0.66±0.07	0.74±0.02
	MVBM-LSTM	0.67±0.03	0.66±0.03	0.74±0.05	0.83±0.03
	US-MLP	**0.81±0.04**	**0.80±0.04**	**0.88±0.03**	**0.91±0.04**
MVCBMs	seq-avg	0.60±0.04	0.55±0.05	0.62±0.06	0.71±0.07
	seq-LSTM	*0.71±0.04*	*0.70±0.04*	*0.75±0.03*	*0.83±0.02*
	joint-avg	0.60±0.12	0.51±0.16	0.62±0.08	0.72±0.06
	joint-LSTM	0.69±0.05	0.67±0.05	0.70±0.05	0.78±0.04
Intervened MVCBMs	seq-avg	0.73±0.05	0.73±0.05	0.82±0.02	0.87±0.02
	seq-LSTM	*0.76±0.04*	*0.74±0.06*	*0.84±0.05*	0.88±0.06
	joint-avg	0.49±0.10	0.39±0.11	0.66±0.14	0.75±0.09
	joint-LSTM	0.76±0.05	0.71±0.06	0.82±0.02	*0.89±0.03*

Table 2. Model performance comparison for predicting the diagnosis over five cross-validation folds on the appendicitis dataset. Metrics are reported as averages followed by standard deviations. For the intervened MVCBMs, interventions were performed on all concepts. *Blue italics* indicates the best result in a group of models, while **bold** indicates the best result overall.

Metric	Model	Concept										
		$\underline{c_1}$	c_2	c_3	$\underline{c_4}$	$\underline{c_5}$	c_6	c_7	$\underline{c_8}$	$\underline{c_9}$	c_{10}	c_{11}
AUROC	Random	0.50	0.50	0.50	0.50	0.50	0.50	**0.50**	0.50	0.50	0.50	**0.50**
	seq-avg	0.58	0.50	0.48	**0.69**	0.64	**0.72**	0.34	*0.55*	0.59	0.51	0.36
	seq-LSTM	**0.73**	*0.56*	*0.56*	*0.68*	**0.85**	*0.67*	*0.44*	**0.63**	**0.63**	**0.64**	0.29
	joint-avg	0.67	0.47	**0.62**	*0.68*	0.62	*0.67*	0.40	0.53	0.57	0.48	0.38
	joint-LSTM	*0.69*	**0.57**	0.49	0.64	*0.83*	0.48	*0.44*	**0.63**	*0.62*	*0.62*	*0.39*
AUPR	Random	0.36	0.46	0.09	0.05	0.76	0.29	*0.07*	0.16	0.15	0.17	*0.03*
	seq-avg	0.43	0.46	*0.11*	**0.21**	*0.82*	**0.55**	0.06	**0.30**	0.20	0.17	**0.04**
	seq-LSTM	**0.60**	*0.53*	*0.11*	*0.17*	**0.94**	*0.50*	**0.08**	*0.26*	*0.24*	**0.30**	*0.03*
	joint-avg	*0.54*	0.43	**0.13**	0.09	0.81	0.48	0.06	0.22	0.19	0.16	*0.03*
	joint-LSTM	*0.54*	**0.55**	0.09	0.07	**0.94**	0.32	0.06	0.24	**0.25**	*0.24*	*0.03*

Table 3. Model performance comparison for concept prediction on the appendicitis dataset. Metrics are reported as averages over five cross-validation folds. **Bold** indicates the best result; *italics* indicates the second best. Concepts for which all models, on average, performed better than the random guess (w.r.t. both AUROC and AUPR) are underlined. Namely, these are *surrounding tissue reaction* (c_1), *visibility of the appendix* (c_4), *pathological lymph nodes* (c_5), *meteorism* (c_8), and *irregular appendix layers* (c_9). For the explanation of the other concepts, see Table 6 in Appendix C; extended results can be found in Table 7, Appendix D.

Intervention Experiment Last but not least, we conducted several experiments to investigate whether intervening on a subset of concepts within an MVCBM model enhances its predictive performance. The figures reported in Table 2 for the intervened MVCBMs were obtained by intervening on *all* concepts, i.e. $\mathcal{S} = \{1, \ldots, K\}$ (Equation 6). Note that interventions improve the performance for all four MVCBM configurations except the jointly trained MVCBM-avg, which is the worst model w.r.t. the largest number of concepts when comparing AUROCs and AUPRs (Table 3). Moreover, the intervened MVCBM with the LSTM-based fusion outperforms all baselines but US-MLP. The right panel of Figure 4 summarises the intervention impact for jointly and sequentially trained MVCBM-LSTM under $\mathcal{S} = \{1, \ldots, K\}$. It depicts how often intervening on the model corrects, spoils, or does not alter the target prediction. Interestingly, when the intervention flips the prediction from false to true, it usually happens for false negatives suggesting that interventions are particularly helpful for detecting the positive class, i.e. the presence of appendicitis. However, to imitate a more realistic scenario where only a subset of true concept values is available at test time, we iteratively select a random subset $\mathcal{S}$ of size ranging from 0 (no intervention) to K in increments of 1 and intervene on the corresponding concept nodes. Cross-validated results of this experiment for LSTM-based MVCBMs are depicted in the left panel of Figure 4. We observe a decreasing tendency in the binary cross-entropy (BCE) loss with the increase in the cardinality of $\mathcal{S}$ for both optimisation procedures.

6. Discussion

The introduced extension of the concept bottleneck models (Koh et al., 2020) to the multiview setting makes them

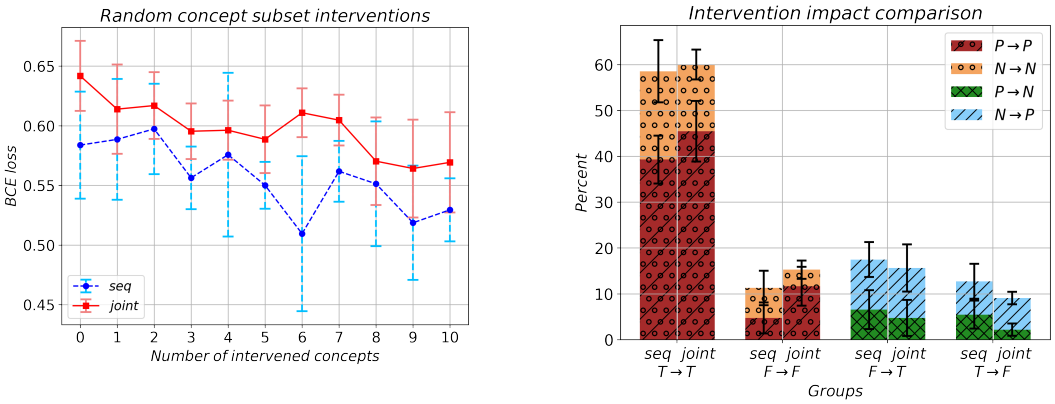

Figure 4. Left: The effect of increasing the number of interventions on a randomly chosen concept subset on the binary cross-entropy loss function for sequentially and jointly trained MVCBM-LSTM aggregated over five cross-validation folds. *Right*: Intervention impact stratified by the label (positive/negative) and correctness (true/false) for sequentially and jointly trained MVCBM-LSTM aggregated over five cross-validation folds.

more readily applicable to medical imaging datasets where multiple images or even modalities may be observable for each subject. In this work, we proposed a practical architecture based on the hybrid fusion approach (Baltrusaitis et al., 2019) that can handle varying numbers of views per data point, partial observability of the concepts from images, and leverage spatial or temporal ordering. To the best of our knowledge, the considered setting has not been explored in the literature despite its relevance to biomedical applications (Zhang et al., 2018; Wang et al., 2020a; Qian et al., 2021).

We demonstrated the feasibility of our model and the benefits of the multiview approach on a natural image benchmarking dataset for attribute-based classification (Section 5.1). Moreover, we applied the model to predict appendicitis in pediatric patients with abdominal pain based on the US data (Section 5.2). Our results suggest that the presented multiview fusion approach is effective and that multiview concept bottlenecks can achieve performance on par with black-box models while allowing medical practitioners to interpret and intervene on the predictions. Most of the prior work on leveraging ML for appendicitis has focused on tabular datasets with handcrafted features (Hsieh et al., 2011; Reismann et al., 2019; Aydin et al., 2020; Marcinkevics et al., 2021) or more invasive imaging modalities (Rajpurkar et al., 2020). This work makes the first step towards the computer-aided diagnosis of appendicitis based on abdominal ultrasound, a noninvasive, accessible, and cheap modality.

Limitations The MVCBM model and experimental setup have several limitations. Similar to the vanilla CBMs, multiview bottlenecks assume a sufficient set of concepts and do not allow for unsupervised representation learning. In practical use cases, this assumption might be restrictive and may prevent the model from achieving a superhuman performance level at the downstream task. The appendicitis dataset (Section 4.1) used in our experiments represents a small, relatively homogeneous patient cohort recruited from a single clinical centre over a short period. Moreover, we did not have histologically confirmed diagnoses in conservatively treated patients. Therefore, further investigation is warranted to conduct external validation and explore failure modes. Last but not least, the current image preprocessing (Section 4.2) discards scale information and makes it impossible to detect the appendix diameter, a relevant sonographic sign of appendicitis (Reddan et al., 2016).

7. Conclusion

Motivated by the demand for model interpretability in biomedical applications, we investigated the use of concept bottleneck models for predicting pediatric appendicitis based on abdominal ultrasound images. We proposed a multiview concept bottleneck model — an extension of the conventional approach to concept-based classification, capable of handling multiple and varying numbers of views of the same object of interest. Our experimental results suggest that MVCBM achieves competitive performance, while also offering an alternative to black-box deep learning models and lending itself to real-time interaction with the end-user.

Future Work We plan to apply our model to an extended cohort consisting of patients recruited between 2016 and 2021. In addition to the prediction of the diagnosis, we also plan to investigate the treatment assignment and disease severity classification. Various model design alterations, such as other choices of learnable fusion, the introduction of unsupervised concepts in the bottleneck layer, or uncertainty quantification, are to be considered as well.

Acknowledgements

RM was supported by the SNSF grant #320038189096, EO was supported by the Hasler Foundation grant #21050.

References

Akmese, O. F., Dogan, G., Kor, H., Erbay, H., and Demir, E. The use of machine learning approaches for the diagnosis of acute appendicitis. *Emergency Medicine International*, 2020:1–8, 2020.

Aydin, E., Türkmen, İ. U., Namli, G., Öztürk, Ç., Esen, A. B., Eray, Y. N., Eroğlu, E., and Akova, F. A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children. *Pediatric Surgery International*, 36(6):735–742, 2020.

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019.

Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Cheplygina, V. Cats or CAT scans: Transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering*, 9:21–27, 2019.

Deleger, L., Brodzinski, H., Zhai, H., Li, Q., Lingren, T., Kirkendall, E. S., Alessandrini, E., and Solti, I. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. *Journal of the American Medical Informatics Association*, 20(e2):e212–e220, 2013.

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

Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. HeMIS: Hetero-modal image segmentation, 2016. arXiv:1607.05194.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Hsieh, C.-H., Lu, R.-H., Lee, N.-H., Chiu, W.-T., Hsu, M.-H., and Li, Y.-C. 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, 2011.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2668–2677. PMLR, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5338–5348. PMLR, 2020.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372. IEEE, 2009.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

Lockhart, J., Marchesotti, N., Magazzeni, D., and Veloso, M. Towards learning to explain with concept bottleneck models: mitigating information leakage, 2022. Workshop on Socially Responsible Machine Learning (SRML), colocated with ICLR 2022.

Losch, M., Fritz, M., and Schiele, B. Semantic bottlenecks: Quantifying and improving inspectability of deep representations. *International Journal on Computer Vision*, 129(11):3136–3153, 2021.

Ma, C., Guo, Y., Yang, J., and An, W. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. *IEEE Transactions on Multimedia*, 21(5):1169–1182, 2019.

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

Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., and Vogt, J. E. Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. *Frontiers in Pediatrics*, 9, 2021.

Marcos, D., Fong, R., Lobry, S., Flamary, R., Courty, N., and Tuia, D. Contextual semantic interpretability. In *Computer Vision – ACCV 2020*, pp. 351–368. Springer International Publishing, 2021.

Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., and Weller, A. Do concept bottleneck models learn as intended?, 2021. arXiv:2105.04289.

Mostbeck, G., Adam, E. J., Nielsen, M. B., Claudon, M., Clevert, D., Nicolau, C., Nyhsen, C., and Owens, C. M. How to diagnose acute appendicitis: ultrasound first. *Insights into Imaging*, 7(2):255–263, 2016.

Nguyen, N. D. and Wang, D. Multiview learning for understanding functional multiomics. *PLoS Computational Biology*, 16(4):e1007677, 2020.

Nie, F., Cai, G., Li, J., and Li, X. Auto-weighted multiview learning for image clustering and semi-supervised classification. *IEEE Transactions on Image Processing*, 27(3):1501–1511, 2018.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch, 2017. NIPS 2017 Autodiff Workshop.

Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., Sun, Q., Lu, L., and Shung, K. K. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nature Biomedical Engineering*, 5(6):522–532, 2021.

Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C. P., Lungren, M. P., Ng, A. Y., and Patel, B. N. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Scientific Reports*, 10(1), 2020.

Reddan, T., Corness, J., Mengersen, K., and Harden, F. Ultrasound of paediatric appendicitis and its secondary sonographic signs: providing a more meaningful finding. *Journal of medical radiation sciences*, 63(1):59–66, 2016.

Reismann, J., Romualdi, A., Kiss, N., Minderjahn, M. I., Kallarackal, J., Schad, M., and Reismann, M. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. *PLoS ONE*, 14(9):e0222030, 2019.

Roig Aparicio, P., Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., and Vogt, J. E. Learning medical risk scores for pediatric appendicitis. In *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Sawada, Y. and Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10: 41758–41765, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Stiel, C., Elrod, J., Klinke, M., Herrmann, J., Junge, C.-M., Ghadban, T., Reinshagen, K., and Boettcher, M. The modified Heidelberg and the AI appendicitis score are superior to current scores in predicting appendicitis in children: A two-center cohort study. *Frontiers in Pediatrics*, 8, 2020.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2021.

van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J. W. L. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

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

Wang, Y., Yue, W., Li, X., Liu, S., Guo, L., Xu, H., Zhang, H., and Yang, G. Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images. *IEEE Access*, 8:52010–52017, 2020b.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD birds 200, 2010. Computation & Neural Systems Technical Report, California Institute of Technology.

Wier, L. M., Yu, H., Owens, P. L., and Washington, R. Overview of children in the emergency department, 2010: Statistical brief #157. In *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*. Agency for Healthcare Research and Quality (US), Rockville (MD), 2013.

Xia, J., Wang, Z., Yang, D., Li, R., Liang, G., Chen, H., Heidari, A. A., Turabieh, H., Mafarja, M., and Pan, Z. Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. *Computers in Biology and Medicine*, 143:105206, 2022.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019.

Xu, C., Tao, D., and Xu, C. A survey on multi-view learning, 2013. arXiv:1304.5634.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhang, C., Adeli, E., Zhou, T., Chen, X., and Shen, D. Multi-layer multi-view classification for Alzheimer's disease diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A. Further Implementation Details

Architectures Table 4 provides a detailed description of the MVCBM's architecture (Section 3) as implemented in our experiments (Section 5). Herein, B denotes the batch size, and K is the number of concepts. Note that, in the appendicitis dataset, all US image sequences have been padded to the length of 20. However, as intended, fusion layers discard the padding and are applied to variable-length sequences. For the MVAwA dataset (Section 4.1), we used a similar architecture. However, we did not pad the sequences (all data points had the same number of views), the input images were 224×224 px^2, and the output layer was 50 units wide and had `Softmax` activation.

Module	Layers	Input Dimensions	Output Dimensions
$\boldsymbol{h}_{\boldsymbol{\psi}}(\cdot)$	`ResNet-18`	(B, 20, 3, 400, 400)	(B, 20, 512)
$\boldsymbol{r}_{\boldsymbol{\xi}}(\cdot)$	`LSTM` or `mean`	(B, 20, 512)	(B, 512)
$\boldsymbol{s}_{\boldsymbol{\zeta}}(\cdot)$	`Linear`	(B, 512)	(B, 256)
	`ReLu`	(B, 256)	(B, 256)
	`Linear`	(B, 256)	(B, 64)
	`ReLu`	(B, 64)	(B, 64)
	`Linear`	(B, 64)	(B, K)
	`Sigmoid`	(B, K)	(B, K)
$f_{\boldsymbol{\theta}}(\cdot)$	`Linear`	(B, K)	(B, 6)
	`ReLu`	(B, 6)	(B, 6)
	`Linear`	(B, 6)	(B, 1)
	`Sigmoid`	(B, 1)	(B, 1)

Table 4. Summary of the MVCBM architecture used for the appendicitis dataset. Here, B denotes the batch size, and K is the number of concepts. A similar architecture was employed in the MVAwA experiments.

Training and Hyperparameters In all experiments, deep learning models were trained using the Adam optimiser (Kingma & Ba, 2015). To avoid potential overfitting on the currently small appendicitis dataset, throughout training, we applied on-the-fly data augmentation with Gaussian noise addition, random black rectangle insertion and one additional randomly chosen transformation from the following list: brightness adjustment, rotation, shearing, resizing, change of image sharpness, or image gamma correction. Applicable model hyperparameter values used for appendicitis and MVAwA datasets are provided in Table 5. There, by E_t and η_t, we denote the number of epochs used to train a model and the initial learning rate, respectively. Note that sequentially trained MVCBMs allow for a separate hyperparameter configuration for the concept model $\boldsymbol{g_\phi}(\cdot)$. We exploit this possibility for the number of epochs (E_c) and the initial learning rate (η_c). Due to the lack of this freedom, we have found that jointly trained MVCBMs require careful selection of E_t and η_t for the model weights to converge. *LR dec. freq.* and *LR dec. fact.* denote how often, w.r.t. the number of epochs, the learning rate is decreased and the factor by which it is decreased, respectively. Finally, recall that parameter α controls the trade-off between the target and concepts loss terms in the jointly trained concept bottleneck models (Equation 5).

(a) MVAwA

Model	**Hyperparameter**							
	E_c	E_t	η_c	η_t	B	LR Dec. Freq.	LR Dec. Fact.	α
Single-CBM-seq	20	20	1.0e-4	1.0e-3	64	30	2	—
Single-CBM-joint	—	50	—	1.0e-4	64	30	2	1.0
MVBM-avg	—	50	—	1.0e-4	64	30	2	—
MVBM-LSTM	—	50	—	1.0e-4	64	30	2	—
MVCBM-seq-avg	20	10	1.0e-4	1.0e-3	64	30	2	—
MVCBM-seq-LSTM	20	10	1.0e-4	1.0e-3	64	30	2	—
MVCBM-joint-avg	—	50	—	1.0e-4	64	30	2	1.0
MVCBM-joint-LSTM	—	50	—	1.0e-4	64	30	2	1.0

(b) Appendicitis

Model	**Hyperparameter**							
	E_c	E_t	η_c	η_t	B	LR Dec. Freq.	LR Dec. Fact.	α
ResNet-18	—	30	—	1.0e-3	4	10	2	—
MVBM-avg	—	100	—	1.0e-4	4	50	2	—
MVBM-LSTM	—	50	—	1.0e-4	4	100	2	—
US-MLP	—	20	—	1.0e-2	4	10	2	—
MVCBM-seq-avg	20	20	1.0e-4	1.0e-2	4	10	2	—
MVCBM-seq-LSTM	20	20	1.0e-4	1.0e-2	4	10	2	—
MVCBM-joint-avg	—	70	—	1.0e-4	4	50	2	1.0
MVCBM-joint-LSTM	—	40	—	1.0e-3	4	100	2	1.0

Table 5. Hyperparameter values of all the models trained on the (a) MVAwA and (b) appendicitis data. Herein, E_c and E_t are the numbers of training epochs for the concept and target models, respectively; η_c and η_t denote the initial learning rates for concept and target models, respectively. In addition, we report the learning rate (LR) decrease frequency (dec. freq.) and decrease factor (dec. fact.).

B. MVAwA Dataset

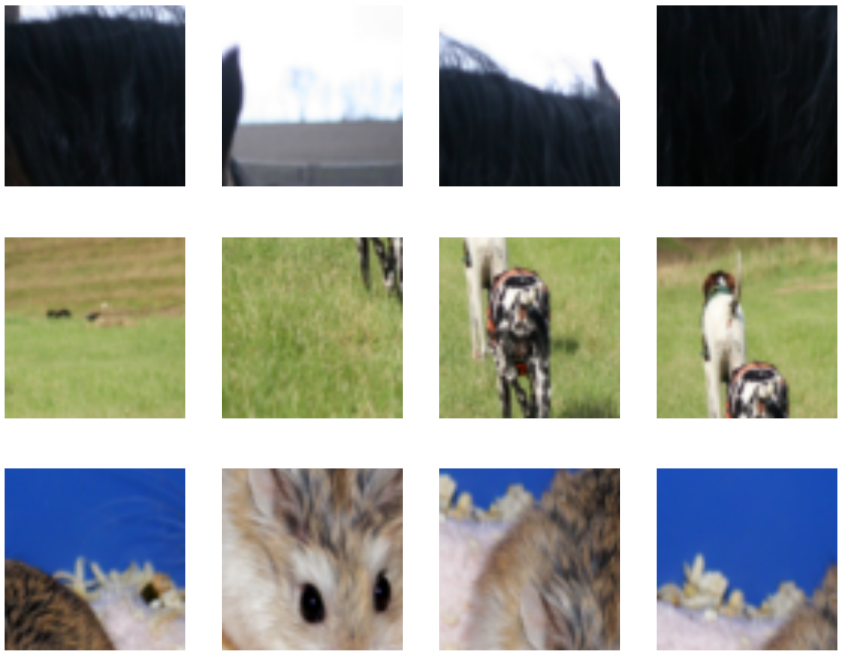

Figure 5. Three examples of the four-view data points from the multiview AwA dataset (Section 4.1); each row corresponds to a single data point. Observe that every view (columns) constitutes a random patch of the original AwA image. Hence, in this dataset, the views are exchangeable. Moreover, note that some concepts can be identified only from certain views, e.g., in the bottom row, attributes referring to the background cannot be detected from the second (counting from the left) view.

C. Appendicitis Dataset

Concept	Explanation	Pos., %	Neg., %
c_1	Surrounding tissue reaction	36	64
c_2	Free intraperitoneal fluid	46	54
c_3	Thickening of the bowel wall	9	91
c_4	Enteritis	5	95
c_5	Visibility of the appendix	76	24
c_6	Pathological lymph nodes	29	71
c_7	Coprostasis	7	93
c_8	Meteorism	16	84
c_9	Irregular appendix layers	15	85
c_{10}	Target sign	17	83
c_{11}	Gynaecological findings	3	97

Table 6. Variables used as concepts in the experiments on the appendicitis dataset. Last two columns report frequencies of the positive (pos.) and negative (neg.) values. As can be seen, some concepts are particularly sparse, e.g. c_{11} (*gynaecological findings*).

D. Further Results

Below, we provide further empirical results obtained on the appendicitis data (Section 5.2). Figure 6 contains the learning curves for the sequentially trained LSTM-based MVCBM in terms of the test binary cross-entropy loss. Table 7 is an extended version of Table 3 from Section 5.2 and contains accuracy, macro-averaged F1 scores, AUROCs, and AUPRs for concept prediction in the appendicitis data.

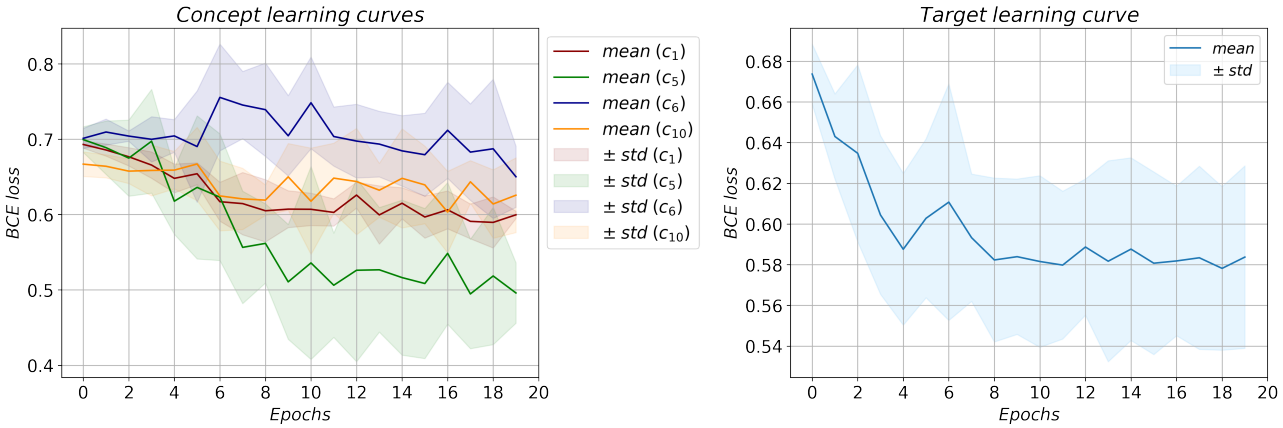

Figure 6. Left: Concept learning curves for the top four best learnt concepts (w.r.t. to the macro-averaged F1 score, AUROC, and AUPR, cf. Table 7) for the sequentially trained LSTM-based model (the best-performing MVCBM configuration, cf. Table 2) on validation data over five cross-validation folds. Concept meanings can be found in Table 6. *Right*: Target variable learning curve for the same model on validation data over five cross-validation folds.

Metric	Model	Concept										
		$\underline{c_1}$	c_2	c_3	$\underline{c_4}$	$\underline{c_5}$	c_6	c_7	$\underline{c_8}$	$\underline{c_9}$	c_{10}	c_{11}
ACC	Random	0.50	0.50	0.50	0.50	0.50	0.50	*0.50*	0.50	0.50	0.50	*0.50*
	seq-avg	0.58±0.10	0.50±0.02	**0.65±0.14**	*0.70±0.16*	0.60±0.09	0.65±0.12	0.33±0.22	0.59±0.15	0.62±0.19	**0.66±0.11**	0.36±0.13
	seq-LSTM	**0.69±0.04**	*0.53±0.04*	0.62±0.10	0.64±0.09	*0.80±0.02*	*0.66±0.07*	0.46±0.16	*0.59±0.05*	*0.62±0.10*	0.60±0.09	0.48±0.19
	joint-avg	0.63±0.06	0.50±0.05	*0.63±0.12*	**0.89±0.08**	0.66±0.15	**0.67±0.05**	**0.58±0.13**	**0.62±0.15**	0.60±0.10	0.52±0.09	**0.73±0.15**
	joint-LSTM	*0.65±0.05*	**0.54±0.08**	0.58±0.16	0.55±0.13	**0.82±0.03**	0.35±0.13	0.37±0.14	0.52±0.08	**0.68±0.13**	*0.63±0.08*	0.38±0.13
Macro F1	Random	0.49	0.50	0.40	0.37	0.46	0.48	**0.39**	0.43	0.43	0.44	*0.36*
	seq-avg	0.49±0.09	0.47±0.02	*0.46±0.09*	0.46±0.08	0.56±0.07	0.60±0.09	0.25±0.13	*0.48±0.08*	0.46±0.11	0.47±0.05	0.28±0.06
	seq-LSTM	**0.66±0.04**	*0.52±0.04*	**0.46±0.06**	*0.46±0.06*	*0.76±0.02*	**0.62±0.08**	0.35±0.10	**0.50±0.05**	**0.52±0.08**	**0.53±0.04**	0.33±0.07
	joint-avg	0.61±0.08	0.50±0.05	**0.46±0.06**	**0.52±0.08**	0.59±0.13	*0.60±0.06*	*0.39±0.05*	*0.48±0.08*	0.47±0.05	0.42±0.06	**0.44±0.07**
	joint-LSTM	*0.61±0.03*	**0.54±0.07**	0.42±0.09	0.40±0.06	**0.77±0.02**	0.34±0.13	0.32±0.10	0.45±0.04	*0.51±0.06*	*0.51±0.06*	0.28±0.07
AUROC	Random	0.50	0.50	0.50	0.50	0.50	0.50	**0.50**	0.50	0.50	0.50	**0.50**
	seq-avg	0.58±0.10	0.50±0.04	0.48±0.12	**0.69±0.13**	0.64±0.07	**0.72±0.08**	0.34±0.10	0.55±0.05	0.59±0.10	0.51±0.04	0.36±0.18
	seq-LSTM	**0.73±0.02**	*0.56±0.07*	*0.56±0.11*	0.68±0.15	**0.85±0.04**	0.67±0.10	0.44±0.15	**0.63±0.12**	**0.63±0.09**	**0.64±0.10**	0.29±0.21
	joint-avg	0.67±0.12	0.47±0.05	**0.62±0.14**	*0.68±0.13*	0.62±0.11	*0.67±0.04*	0.40±0.15	0.53±0.09	0.57±0.12	0.48±0.09	0.38±0.22
	joint-LSTM	*0.69±0.07*	**0.57±0.08**	0.49±0.12	0.64±0.06	*0.83±0.06*	0.48±0.09	*0.44±0.10*	*0.63±0.15*	*0.62±0.07*	*0.62±0.08*	*0.39±0.22*
AUPR	Random	0.36	0.46	0.09	0.05	0.76	0.29	*0.07*	0.16	0.15	0.17	*0.03*
	seq-avg	0.43±0.08	0.46±0.09	*0.11±0.06*	**0.21±0.19**	0.82±0.06	**0.55±0.20**	0.06±0.03	**0.30±0.08**	0.20±0.07	0.17±0.04	**0.04±0.03**
	seq-LSTM	**0.60±0.10**	*0.53±0.08*	*0.11±0.06*	*0.17±0.17*	**0.94±0.02**	*0.50±0.22*	**0.08±0.05**	*0.26±0.08*	*0.24±0.10*	**0.30±0.13**	0.03±0.03
	joint-avg	0.54±0.13	0.43±0.08	**0.13±0.03**	0.09±0.05	0.81±0.07	0.48±0.18	0.06±0.04	0.22±0.06	0.19±0.09	0.16±0.05	0.03±0.02
	joint-LSTM	*0.54±0.10*	**0.55±0.03**	0.09±0.03	0.07±0.02	*0.94±0.03*	0.32±0.17	0.06±0.03	0.24±0.09	**0.25±0.11**	*0.24±0.09*	0.03±0.02

Table 7. Model performance comparison for concept prediction over five cross-validation folds on the appendicitis dataset (this is an extended version of Table 3 from the main text). Metrics are reported as averages followed by standard deviations. **Bold** indicates the best result; *italics* indicates the second best. Concepts for which all models, on average, performed better than the random guess (w.r.t. all metrics) are underlined. Namely, these are *surrounding tissue reaction* (c_1), *visibility of the appendix* (c_4), *pathological lymph nodes* (c_5), *meteorism* (c_8), and *irregular appendix layers* (c_9). For the explanation of concepts, see Table 6 in Appendix C.