Image quality assessment for machine learning tasks using meta-reinforcement learning

Shaheer U. Saeed [a,*], Yunguan Fu [a,b], Vasilis Stavrinides [c,d], Zachary M.C. Baum [a], Qianye Yang [a], Mirabela Rusu [e], Richard E. Fan [f], Geoffrey A. Sonn [e,f], J. Alison Noble [g], Dean C. Barratt [a], Yipeng Hu [a,g]

[a] Centre for Medical Image Computing, Wellcome/EPSRC Centre for Interventional & Surgical Sciences, and Department of Medical Physics & Biomedical Engineering, University College London, London, UK
[b] InstaDeep, London, UK
[c] Division of Surgery & Interventional Science, University College London, London, UK
[d] Department of Urology, University College Hospital NHS Foundation Trust, London, UK
[e] Department of Radiology, Stanford University, Stanford, California, USA
[f] Department of Urology, Stanford University, Stanford, California, USA
[g] Department of Engineering Science, University of Oxford, Oxford, UK

ARTICLE INFO

ABSTRACT

In this paper, we consider image quality assessment (IQA) as a measure of how images are amenable with respect to a given downstream task, or *task amenability*. When the task is performed using machine learning algorithms, such as a neural-network-based *task predictor* for image classification or segmentation, the performance of the task predictor provides an objective estimate of task amenability. In this work, we use an *IQA controller* to predict the task amenability which, itself being parameterised by neural networks, can be trained simultaneously with the task predictor. We further develop a meta-reinforcement learning framework to improve the adaptability for both IQA controllers and task predictors, such that they can be fine-tuned efficiently on new datasets or meta-tasks. We demonstrate the efficacy of the proposed task-specific, adaptable IQA approach, using two clinical applications for ultrasound-guided prostate intervention and pneumonia detection on X-ray images.

1. Introduction

1.1. Image quality assessment

Medical imaging is used extensively for diagnostic and therapeutic procedures in medicine, whether they be interventional or non-interventional in nature. Several such diagnostic, navigational or therapeutic tasks in the clinical workflow rely on medical images where they inform the clinician's judgement, directly or via derived measurements. Medical imaging is increasingly being used as a navigational aid to guide surgical and other interventional procedures, such as for prostate biopsies (Brown et al., 2015), liver resections (Simpson and Kingham, 2016), and brain resections (Kondziolka and Lunsford, 1996). Treatment planning, for example radiotherapy planning, relies heavily on pre-operative medical images (Dirix et al., 2014; Liney and Moerland, 2014). Moreover, imaging is commonly used for diagnostic clinical tasks whether the task is performed manually by humans or automated using computer aided diagnosis. The use of chest radio-graphs for diagnosis of lung diseases (Doi, 2007), computed tomography (CT) or magnetic resonance (MR) scans for diagnosis of brain diseases (Doi, 2007) and ultrasound (US) for diagnosis of uterine diseases (Dueholm, 2006), are all common examples where diagnosis relies heavily or solely on medical images.

The performance of clinical tasks that rely on medical imaging can be adversely impacted by the image quality of the images being used (Chow and Paramesran, 2016). Image quality assessment (IQA) is an effective way to ensure that any clinical task intended for a medical image can be performed reliably as the use of poor quality images for clinical tasks can result in inaccurate,

---

* Corresponding author: .
  E-mail address: shaheer.saeed.17@ucl.ac.uk (S.U. Saeed).

or potentially erroneous, diagnoses or measurements (Davis et al., 2009; Wu et al., 2017; Chow and Paramesran, 2016). IQA serves as a mechanism to ensure that intended downstream target tasks for medical images, such as diagnostic, therapeutic or navigational tasks, can be performed effectively and reliably.

The use of IQA in medicine to ensure reliability in clinical task performance is common and various approaches to IQA have been proposed in the past decades to address this problem. Existing IQA methods broadly fall into two categories, manual and automated assessment. Manual assessment is widely used in clinical practice (Chow and Paramesran, 2016), often involving human interpretation of a set of criteria in order to assess image quality (Loizou et al., 2006; Hemmsen et al., 2010; Shima et al., 2007). Due to the high variance in manual assessment, consensus or mean quality scores from multiple observers may be used to assess images (De Angelis et al., 2007; Chow and Rajagopal, 2015). Although consensus-based methods are able to reduce the variance in predictions and produce repeatable measurements for IQA, the human cost associated with obtaining IQA scoring from multiple expert observers is high (Chow and Paramesran, 2016). Automated assessment methods for IQA provide a means to ensure reproducible measurements and to reduce both the variance in predictions and the involvement of human perception of both the medical image and the IQA criteria (Chow and Paramesran, 2016). These automated methods differ from manual methods in that after a computational tool construction phase, such as training a machine learning model, they are able to quantify IQA without requiring human judgement for any new images being assessed. It is, however, important to note that model construction itself may require human judgement, possibly in the form of labelled samples or to select common features across high or low quality images, for both development and validation of the model.

Automated methods can further be classified based on the extent to which they utilise information from a reference set of images in the IQA model construction phase (Chow and Paramesran, 2016). Full-reference automated methods use selected reference images directly in order to compute an IQA metric, which is often based on a similarity measure between the image being assessed and the subjectively selected reference standard good-quality image. The selection of the reference standard and construction of the metric may be considered as part of model construction (Fuderer, 1988; Kaufman et al., 1989; Henkelman, 1985; Dietrich et al., 2007; Shiao et al., 2007; Geissler et al., 2007; Salem et al., 2002; Choong et al., 2006; Daly, 1992; Jiang et al., 2007; Miao et al., 2008; Wang et al., 2004; Kumar and Rattan, 2012; Kumar et al., 2011; Rangaraju et al., 2012; Kowalik-Urbaniak et al., 2014; Huo et al., 2006). These full-reference methods and other reduced-reference methods, which use partial information from a selected reference image set (Chow and Paramesran, 2016), can thus produce automated, reliable and repeatable measurements. No-reference methods aim to eliminate the subjective selection of a reference standard and do not rely on a reference image set for model construction (Dutta et al., 2013; Kalayeh et al., 2013; Mortamet et al., 2009; Woodard and Carley-Spencer, 2006; Eck et al., 2015; Racine et al., 2016; Davis et al., 2009; Khler et al., 2013; Loizou et al., 2006). These methods require robust mathematical models to capture common features across low quality images or the statistics of high quality image generation, and are therefore specific to certain modalities or applications. For example, methods have been proposed for MR images (Mortamet et al., 2009; Woodard and Carley-Spencer, 2006), single photon emission CT (Kalayeh et al., 2013), and CT (Eck et al., 2015; Racine et al., 2016). Constructing these modality- or application- specific models may not require human labels of IQA at model construction or inference, however, it requires in-depth knowledge of the noise sources within the imaging modality as well as of the image acqui-

sition process and that of intended use of the images (Chow and Paramesran, 2016). Learning-based approaches use machine learning methods, including recent deep learning, to automate IQA and provide fast inference. Most of the existing methods learn from subjective expert labels of IQA (Wu et al., 2017; Zago et al., 2018; Esses et al., 2018; Baum et al., 2021; Liao et al., 2019; Abdi et al., 2017; Lin et al., 2019; Camps et al., 2020), based on a set of pre-selected reference images to varying extent. Most automated IQA methods, including full-, reduced- and no- reference methods, both learning- and non-learning based, were validated against human labels of IQA regardless of the extent to which they use human labels for model construction (Chow and Paramesran, 2016).

In this work we turn our attention to task-specific IQA methods, when the goal of an IQA method is to ensure that specific downstream tasks can be reliably performed, and task-agnostic IQA methods may not always be effective, efficient or feasible. We propose to use the term 'task amenability' to define the usefulness of an image for a specific downstream target task. Most IQA methods, regardless of the extent to which they utilise a reference standard, are often designed to be specific to a certain imaging modality or to a specific anatomy, rather than directly quantify the impact on the downstream clinical task performance (Chow and Paramesran, 2016; Wu et al., 2017; Esses et al., 2018; Eck et al., 2015; Racine et al., 2016), as discussed above. For example, a strong ultrasound reflection, obstructing a significant part of gland boundaries, is catastrophic to a downstream task of gland segmentation, but may not adversely impact a classification task for identifying the presence of the gland. Similar examples also include severe noise found outside of regions of interest, upon which the diagnosis decision does not rely.

Perhaps more interestingly, as increasing number of downstream tasks are being automated by, for example, machine learning models, the subjectively perceived task-specific IQA may be very different from the actual impact of an image on the automated target task. Thus the extent to which these human-defined criterion can measure impact on machine-automated task performance is still an open question. Thus one could argue that, where the downstream target task is independently learned or performed, task amenability may be difficult to quantify.

1.2. Related work and contribution

In our previous work (Saeed et al., 2021a), we proposed a method to objectively quantify task amenability for a specific task by jointly learning the task-specific IQA and the target task to capture the inter-dependence between the two functions, without the need for human labels of IQA. We proposed to simultaneously learn a task predictor, which performs a downstream target task, and an IQA controller, which selects or weights images based on their task amenability. In this scenario, images can be selected or weighted based on their task amenability such that this selected or weighted subset results in improved target task performance. It should be noted that only performing the target task (i.e. using a task predictor function with fixed weights) while learning the task amenability may also be able to capture the dependence of the controller on the target task. However, to capture the inter-dependence which arises from the task predictor training data modification by the controller and the controller training based on task predictor performance, requires the target task to be learnt within the framework. Nonetheless, the fixed task predictor formulation may be useful for applications where the task-predictor has been pre-trained with different data or in applications where what the IQA entails does not impact task performance.

In the proposed problem setting, optimising the controller is dependent on the task predictor being optimised. The problem can thus be modelled in a meta-learning framework where the down-

stream target task performance is maximised with respect to the controller selected images. The task predictor and the data used to train it are considered to be contained within an environment which reflects a Markov decision process (MDP).

Meta-learning problems have increasingly been formulated as reinforcement learning (RL) problems under RL-based meta-learning. In the RL-based meta-learning framework, a parameter associated with the target task is modified by the controller such that the target task performance can be maximised. The reward, which indicates how well the target task is performed, is computed after the parameter modification and thus indicates the effect of the controller's modification. This reward is used to update the controller in a way which maximises the cumulative obtained reward and thus allows for a parameter setting that maximises the target task performance, to be learnt. The target task can include any automated classification, regression or segmentation task and the parameter modification by the controller can include selecting a data transformation strategy for augmentation policy search (Cubuk et al., 2019; Zhang et al., 2019), selecting hyper-parameters like filters for convolutional layers for neural architecture search (Zoph and Le, 2017) and sampling training data for data valuation (Yoon et al., 2020). This approach to learn task amenability bears some resemblance with the data valuation framework presented in (Yoon et al., 2020), however, there are several differences in reward formulation without human labels of quality, the RL algorithms used, and other methodological details including use of the controller for holdout data.

In another previous work we proposed a meta-RL training scheme for the controller (Saeed et al., 2021b). The training scheme involved a recurrent neural network (RNN) based controller and sampling different MDP environments, each with different observer labels for the target task, during training in order to equip the trained controller with adaptability to new labelling standards. The resulting adaptability from data labelled by non-expert observers to high-quality expert labelled data, carefully curated by reviewed consensus, proved useful for the efficient use of labelled data.

In this work, we summarise the two preliminary sets of experimental results (Saeed et al., 2021a; 2021b) and present a general framework to learn an adaptable task amenability assessment using meta-reinforcement learning (meta-RL). The proposed scheme samples environments for training, from a distribution of MDP environments, such that the controller can adapt to new environments sampled from the distribution with a few interactions. Adaptive behaviour is learnt as a result of the proposed training scheme which involves sampling multiple environments and the use of a recurrent neural network (RNN) to equip adaptability to the controller similar to the framework proposed in our previous work (Saeed et al., 2021b). However, different from the previous work, the distribution of environments in the general framework can be over different target tasks, imaging modalities, observer labels, task predictor functions (e.g. different network architectures), or any other variable within the environment, as opposed to over different observer labels (where each sampled environment can be considered a new meta-task). Meta-RL allows for the task predictor and controller functions to be trained together to capture the inter-dependence between them and also to equip adaptability to the controller over different meta-tasks. Therefore, the focus of this work is to formulate a general meta-RL framework for IQA which is applicable to a variety of scenarios where adaptability is to be learnt over a distribution of environments or meta-tasks.

Contributions of this work are summarised as follows: 1) we present a literature review with a contextual discussion for positioning the proposed task-specific IQA methodology and explains the need for co-learning between the IQA controller and a task predictor; 2) building on our preliminary work presented in two recent conference papers (Saeed et al., 2021a; 2021b), we formulate a general, unified meta-RL based meta learning framework to train an adaptable IQA system which directly accounts for the inter-dependence between the task-specific IQA and the target task; 3) we present new experiments to evaluate the hyper-parameters and design choices in our proposed framework; 4) in addition to summarising the two previously reported target tasks of prostate classification and segmentation, we also evaluate for a new diagnostic target task of pneumonia detection, on a public dataset of chest X-ray images; 5) the code used in this study are made available for reproducibility of the presented experimental results.

2. Methods

2.1. Problem formulation

In this section, we formulate the image quality assessment problem in a specific scenario, that uses the output of the measure of quality for task amenable data selection. As described in Sect. 1 and illustrated in the Fig. 1, the proposed IQA formulation relies on two inter-dependent functions, the controller and the task predictor.

2.1.1. Task predictor and IQA controller

Without loss of generality, assume the task predictor is a parametric function,

$$f(x; w) : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

which preforms the target task given an image sample $x \in \mathcal{X}$ and outputs a prediction $y \in \mathcal{Y}$, with parameters w. The controller is also a parametric function,

$$h(x; \theta) : \mathcal{X} \to [0, 1], \tag{2}$$

which outputs a task amenability score given an image sample, x, with parameters θ. In this formulation $\mathcal{X}$ and $\mathcal{Y}$ correspond to the image and label domains for the specific downstream target task respectively. $\mathcal{P}_X$ and $\mathcal{P}_{XY}$ denote the image and joint image-label distributions with probability density functions $p(x)$ and $p(x, y)$, respectively.

The task predictor performs the downstream target task and the controller selects or weights data used for training the task predictor. The task performance informs the controller decisions over time in order to allow for task performance to be improved. The training methodology outlined below helps to capture this inter-dependence between the two functions.

2.1.2. Optimising task predictor

Given that a loss function, $L_f : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, measures how well the target task is performed by the task predictor $f(x; w)$ given task label y, the task predictor is optimised by minimising a weighted loss function as follows:

$$\min_w \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}}[L_f(f(x; w), y)h(x; \theta)]. \tag{3}$$

Here, weighting by the controller-measured task amenability for the same image sample, x, ensures that high loss samples with low task amenability should be weighted less. This provides an incentive to reject such samples, and to accept samples with high task amenability, in optimising the controller described as follows.

2.1.3. Optimising controller

The controller is optimised by minimising a weighted metric function on the validation set $L_h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$:

$$\min_\theta \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}}[L_h(f(x; w), y)h(x; \theta)], \tag{4}$$

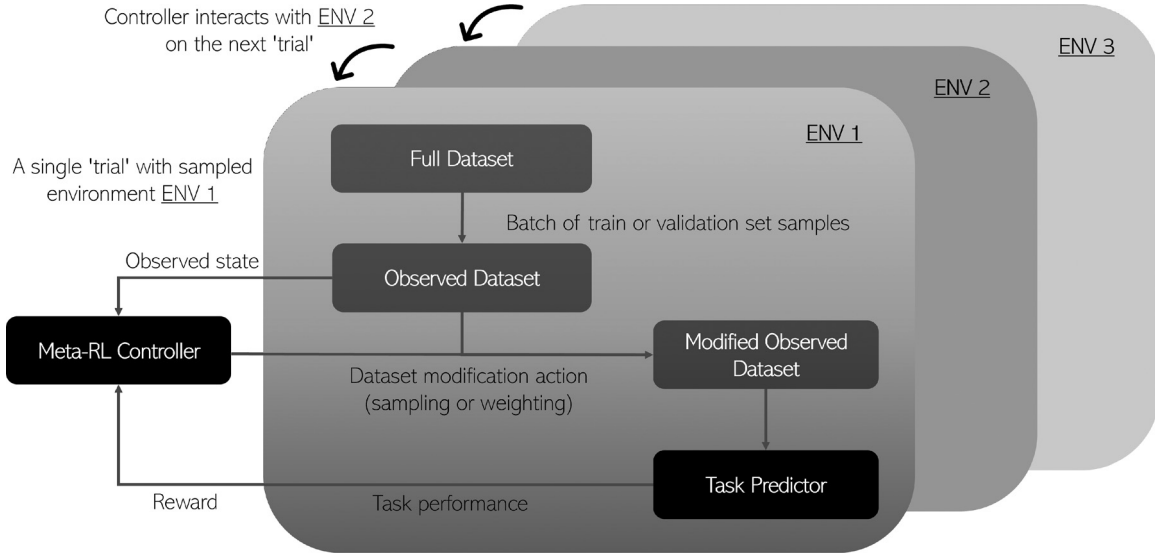**Fig. 1.** Illustration of the proposed multi-environment meta-RL task amenability framework.

s.t. $\mathbb{E}_{x \sim \mathcal{P}_X}[h(x; \theta)] \geq c > 0.$ (5)

The controller thus predicts lower task amenability for samples with higher values from this metric function, which translates to lower task performance, due to the weighted sum being minimised. Intuitively, this means correctly predicting the task labels for lower task amenability samples tends to be difficult. The trivial solution of $h \equiv 0$ is prevented by introducing the constraint.

2.1.4. Bi-level optimisation for learning task amenability

The proposed task amenability framework can thus be posed as the following bi-level minimisation problem (Sinha et al., 2018):

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}}[L_h(f(x; w^*), y)h(x; \theta)],$$ (6)

s.t. $w^* = \arg\min_{w} \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}}[L_f(f(x; w), y)h(x; \theta)],$ (7)

$$\mathbb{E}_{x \sim \mathcal{P}_X}[h(x; \theta)] \geq c > 0.$$ (8)

This problem can be re-structured to permit sampling or selection based on controller outputs by considering the data x and (x, y) to be sampled from the controller-selected or -sampled distributions $\mathcal{P}_X^h$ and $\mathcal{P}_{XY}^h$, with probability density functions $p^h(x) \propto p(x)h(x; \theta)$ and $p^h(x, y) \propto p(x, y)h(x; \theta)$, respectively. Thus, reformulating to facilitate sampling or selection, we can re-write the bi-level minimisation problem as follows:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}^h}[L_h(f(x; w^*), y)],$$ (9)

s.t. $w^* = \arg\min_{w} \mathbb{E}_{(x,y) \sim \mathcal{P}_{XY}^h}[L_f(f(x; w), y)],$ (10)

$$\mathbb{E}_{x \sim \mathcal{P}_X^h}[1] \geq c > 0.$$ (11)

2.2. Meta-reinforcement-learning for task amenability

The formulated task amenability assessment problem, eq. 1 to 9, can be learnt in a RL-based meta-learning framework as formulated in our previous work (Saeed et al., 2021a). In this work we outline a general meta-RL based meta-learning framework to learn adaptable task amenability assessment.

2.2.1. Markov decision process environment construction

The proposed formulation can be modelled as a finite-horizon Markov decision process (MDP) with the controller interacting with, and influencing, an 'environment', which contains the task predictor and the data used to train such a function, as illustrated in Fig. 2. The MDP environment for this task amenability problem thus consists of the data from $\mathcal{P}_{XY}$, where this joint image-label distribution is defined as $\mathcal{P}_{XY} = \mathcal{P}_X \mathcal{P}_{Y|X}$, and the target task predictor $f(\cdot; w)$. At time-step t, the observed state of the environment $s_t = (f(\cdot; w_t), \mathcal{B}_t)$ is composed of the target task predictor $f(\cdot; w)$ and a batch of samples $\mathcal{B}_t = \{(x_i, y_i)\}_{i=1}^{B}$ from a train set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ from the distribution $\mathcal{P}_{XY}$. If each MDP environment is defined as M_k, the joint image-label distribution and task predictor within the environment can be defined as $\mathcal{P}_{XY,k}$ and $f_k(\cdot; w_k)$, respectively. However, in further analysis, we omit k from these expressions, for notational convenience.

2.2.2. Reinforcement learning for bi-level optimisation

Reinforcement learning allows for the training of a controller to maximise a reward obtained based on controller-environment interactions, which are considered to be a MDP. In RL, the MDP is considered to be a 5-tuple $(\mathcal{S}, \mathcal{A}, p, r, \pi)$. $\mathcal{S}$ is the state space and $\mathcal{A}$ is the continuous action space. $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the state transition distribution conditioned on state-actions, e.g. $p(s_{t+1}|s_t, a_t)$ represents the probability of the next state $s_{t+1} \in \mathcal{S}$ given the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$.

The reward function is denoted by $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $R_t = r(s_t, a_t)$ denotes the reward given current state s_t and action a_t. The policy, $\pi(a_t|s_t) : \mathcal{S} \times \mathcal{A} \in [0, 1]$, represents the probability of performing the action a_t given the state s_t. The controller interacting with an environment creates a trajectory of states, actions and rewards, $(s_1, a_1, R_1, s_2, a_2, R_2, \ldots, s_T, a_T, R_T)$, where the subscript indicates the time-step.

The goal of the agent is to maximise the cumulative reward over a trajectory. The cumulative reward is the discounted sum of accumulated rewards starting from time-step t: $Q^\pi(s_t, a_t) = \sum_{k=0}^{T} \gamma^k R_{t+k}$, where the discount factor $\gamma \in [0, 1]$ is used to discount future rewards. The objective of the controller is thus to learn a parameterised policy π_θ which maximises the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[Q^\pi(s_t, a_t)]$. The central optimisation problem in RL can be expressed as:

$$\theta^* = \text{argmax}_\theta J(\theta)$$
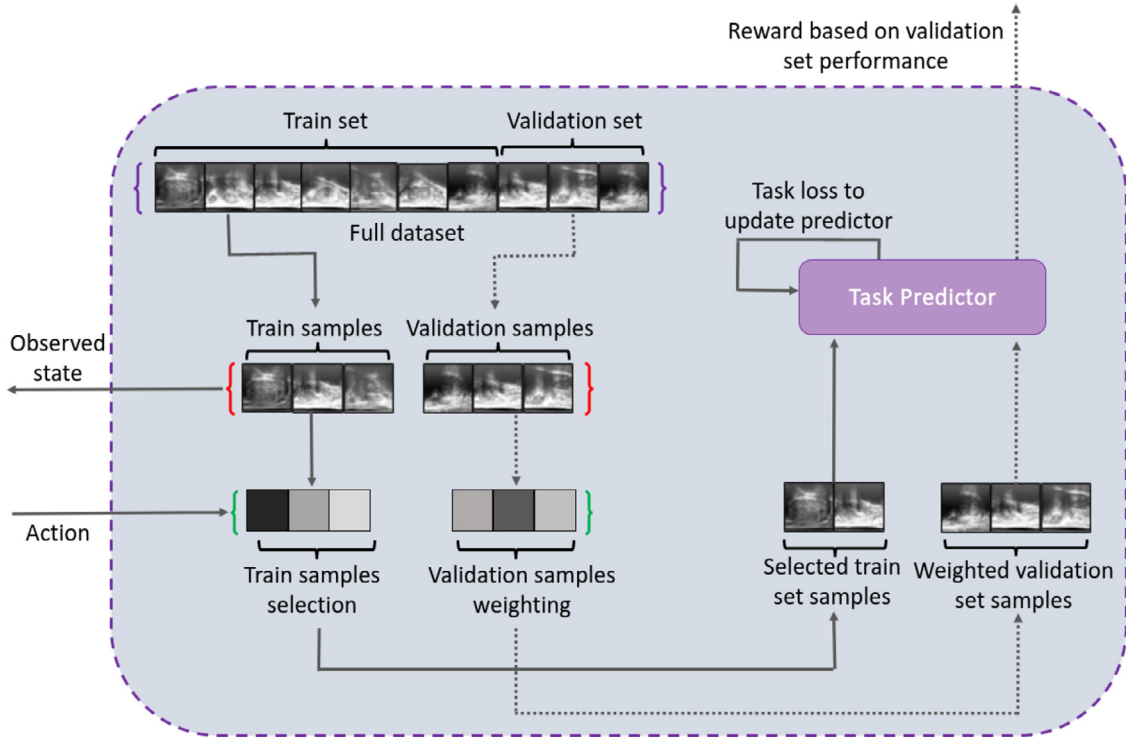
where θ^* denotes optimal policy parameters.

Fig. 2. A single environment in the IQA framework.

In our previous work (Saeed et al., 2021a), we proposed to train the task-amenability-predicting controller using RL, where the controller outputs sampling probabilities $\{h(x_{i,t}, \theta)\}_{i=1}^{B}$ based on the input images. The action $a_t = \{a_{i,t}\}_{i=1}^{B} \in \{0, 1\}^{B}$ leads to a sample selection decision for target task predictor training, if $a_{i,t} = 1$. The selection is done based on $a_{i,t} \sim \text{Bernoulli}(h(x_{i,t}; \theta))$. The policy $\pi_\theta(a_t|s_t)$ is defined as:

$$\log \pi_\theta(a_t|s_t) = \sum_{i=1}^{B} h(x_{i,t}; \theta)a_{i,t} + (1 - h(x_{i,t}; \theta)(1 - a_{i,t})) \qquad (12)$$

In this formulation, the reward R_t was formulated based on the metric function which measures performance of the target task, L_h. There are several options for how to design a metric using L_h and we describe a few examples in the subsequent paragraphs.

2.2.3. Meta reinforcement learning for adaptability

In this work, we propose to train the controller using meta-RL that unifies the single-environment formulation proposed in (Saeed et al., 2021a). Meta-RL is a training procedure which has the same goal of maximising the expected return as RL, however, the objective is averaged across multiple MDP environments in a distribution of MDPs such that the trained controller can effectively generalise or adapt to new MDPs sampled from the distribution (Duan et al., 2016; Wang et al., 2017; Botvinick et al., 2019). To facilitate adaptability in the controller, it is shared across different MDPs sampled from the distribution of MDPs $\mathcal{P}_M$. A period of interaction with a single MDP is referred to as a *trial*. The controller learns across multiple trials by sampling an MDP $M_k \sim \mathcal{P}_M$ for each trial. The controller also takes the action a_t, raw reward r_t, and termination flag d_t at the previous time step in addition to the observed current state s_{t+1}. Note that for per-sample operation $r_t = R_t$ at the episode end, and zero otherwise, similar to sparse reward formulations in (Duan et al., 2016; Wang et al., 2017). Additionally, the controller embeds a recurrent neural network (RNN) with the internal memory shared across episodes within the same

trial. The internal memory is reset before the controller encounters another environment i.e. at the start of each trial. This mechanism allows for adaptability even with fixed wights (Duan et al., 2016; Wang et al., 2017; Botvinick et al., 2019). The additional inputs along with the embedded RNN and its internal memory make the controller a function of the history leading up to a sample such that changing history can influence the action for that sample. It should be noted that each sampled MDP M_k, has its own task predictor and joint image-label distributions. There may be benefit in sharing components between the MDPs and a few such cases are discussed in the following paragraphs. The training scheme is outlined in Algorithm 1 and summarised in Fig. 1.

Subsequent to training using the meta-RL scheme, adaptation can be carried out by sampling a single MDP $M_a \sim \mathcal{P}_M$, where M_a is the environment to be adapted to, resetting the RNN internal state once at the start of the adaptation and allowing for controller-environment interaction across multiple episodes. This means following the same scheme as Algorithm 1 but without sampling a new MDP on each trial, without resetting the RNN internal state at the start of each trial, and without updating the controller. The adaptability in this adaptation scheme is a result of updating RNN internal state rather than updates of the weights.

This meta-RL training and adaptation scheme could potentially be applicable to a wide range of scenarios and could be used to train an IQA system which could adapt to different target tasks, new labels for the target task, different task predictor functions, different reward metrics, or different imaging modalities. This could allow for efficient use of data without the need to retrain an entirely new IQA system. As an example, to train an IQA system adaptable to different task predictor functions, for example neural network architectures, the defined MDP distribution $\mathcal{P}_M$ would have MDPs each with a task predictor with a different architecture. For some of these applications, such as to train an adaptable IQA system across multiple reward metrics or labels for the target task, it may be beneficial to share the task predictor between MDP environments. This equips adaptability to the task predictor in addi-

---

Algorithm 1: Adaptable image task amenability assessment using multiple environments.

Data: Multiple MDPs $M_k \sim \mathcal{P}_M$.
Result: Controller $h(\cdot; \theta)$.

while *not converged* **do**
    Sample an MDP $M_k \sim \mathcal{P}_M$;
    Reset the internal state of controller h;
    for *Each episode in all episodes* **do**
        for $t \leftarrow 1$ **to** T **do**
            Sample a training mini-batch $\mathcal{B}_t = \{(x_{i,t}, y_{i,t})\}_{i=1}^B$;
            Compute selection probabilities
            $\{h_{i,t}\}_{i=1}^B = \{h(\tau_{i,t}; \theta_t)\}_{i=1}^B$;
            Sample actions $a_t = \{a_{i,t}\}_{i=1}^B$ w.r.t.
            $a_{i,t} \sim \text{Bernoulli}(h_{i,t})$;
            Select samples $\mathcal{B}_{t,\text{selected}}$ from $\mathcal{B}_t$;
            Update predictor $f(\cdot; w_t)$ with $\mathcal{B}_{t,\text{selected}}$;
            Compute reward R_t;
         end
        Collect one episode $\{\mathcal{B}_t, a_t, R_t\}_{t=1}^T$;
        Update controller $h(\cdot; \theta)$ using the RL algorithm;
    end
end

---

tion to the controller. Moreover, sharing certain components of the dataset may be useful as well, such as for multiple observer labels, it may be useful to share images across environments and define the distribution of MDPs $\mathcal{P}_M$ to be over different joint image-label distributions as outlined in Sect. 2.4.

2.2.4. Reward formulation

In the previous sections we outlined how a performance metric can be used as the reward function to train the task amenability-predicting controller. This allows for many different combinations and variations in the reward formulation, however, in this work we focus on a performance metric computed on the validation set $\mathcal{D}_{\text{val}} = \{(x_j, y_j)\}_{j=1}^M$, from the same distribution as the train set $\mathcal{P}_{XY}$, as: $\{l_{j,t}\}_{j=1}^M = \{L_h(f(x_j; w_t), y_j)\}$. This performance can be used to formulate the un-clipped reward $\tilde{R}_t$. $\tilde{R}_t$ can be formulated in several ways, with or without the controller output for the validation set $\{h_j\}_{j=1}^M = \{h(x_j; \theta)\}_{j=1}^M$. In this work we consider three formulations for $\tilde{R}_t$:

1. $\tilde{R}_{\text{avg},t} = -\frac{1}{M} \sum_{j=1}^M l_{j,t}$, the average performance.
2. $\tilde{R}_{\text{w},t} = -\frac{1}{M} \sum_{j=1}^M l_{j,t} h_j$, the weighted sum.
3. $\tilde{R}_{\text{sel},t} = -\frac{1}{M'} \sum_{j'=1}^{M'} l_{j',t}$, the average of the selected M' samples.

The first reward formulation $\tilde{R}_{\text{avg},t}$ requires pre-selection of highly task amenable data by a human observer to form the validation set since no controller weighting or sampling is performed. These "task amenability" labels can be acquired in addition to the task labels and can be used to form such a clean fixed validation set. The second $\tilde{R}_{\text{w},t}$ and third $\tilde{R}_{\text{sel},t}$ reward formulations do not require these human labels of task amenability as they utilise the controller output to weight or sample the validation set, respectively. For the third reward formulation, $\tilde{R}_{\text{sel},t}$, $\{j'\}_{j'=1}^{M'} \subseteq \{j\}_{j=1}^M$ and $h_{j'} \le h_{k'}, \forall k' \in \{j'\}^c, \forall j' \in \{j'\}$, i.e. the un-clipped reward $\tilde{R}_{\text{sel},t}$ is the average of $\{l_{j'}\}$ from the subset of $M' = \lfloor (1 - s_{rej})M \rfloor$ samples, by removing $s_{rej} \times 100\%$ samples from the end, after sorting h_j in decreasing order. During training, this $\tilde{R}_t$ value is clipped using a moving average $\bar{R}_t = \alpha_R \bar{R}_{t-1} + (1 - \alpha_R)\tilde{R}_t$, where α_R is a hyper-parameter set to 0.9. It is possible to clip the reward using other values such as using a random selection baseline rather than using

a moving average. It is interesting to note that since the validation set is formed of multiple samples, this means that the controller will not only keep samples on which the target task can be performed perfectly but will try to select samples to facilitate generalisability to new samples. Additionally, since the validation set is either weighted or selective, with or without human labels of IQA, with respect to task amenability, generalisability to samples that have high task amenability is encouraged compared to those with low task amenability. This means that cases such as all samples being selected or all but one samples being rejected, are discouraged in this formulation.

2.3. Single environment cases

The task-specific IQA system presented in (Saeed et al., 2021a) can be considered a special case of the general meta-RL framework presented in this work. In the single environment case, there are no environment-specific trials but rather we can consider the same environment to be sampled at the start of each trial. Moreover, adaptation is not required after training. The Reptile update (introduced in Sect. 2.4) is also not applicable to this single environment case and the RNN may be replaced with a simple feed-forward function such as a deep neural network and the additional inputs of a_t, r_t, and d_t may be omitted. More succinctly, meta-RL simplifies to RL for this single environment case. This is useful when adaptability over a dataset distribution is not required but generalisability to new samples in a dataset is still desirable.

2.4. Multiple environments for inter-observer labels

Our previous work presented preliminary results for such a specific formulation of the meta-RL framework (Saeed et al., 2021b), in which a trained IQA system could adapt across multiple observers of task amenability, for example to use non-expert target task labels for training and then use limited expert-labelled data for adaptation to adapt the task-specific IQA definition to a new reference standard. In this paper, the general framework presented in Sect. 2.2.2 is a more general case of the multi-observer framework presented in (Saeed et al., 2021b). Additional results from a comprehensive set of experiments are also presented in Sect. 4.

For the multi-observer setting we consider multiple label distributions $\{\mathcal{P}_{Y|X}^k\}_{k=1}^K$ which means that each sample x has multiple labels $\{y^k\}_{k=1}^K$. Therefore, we have multiple joint image-label distributions $\mathcal{P}_{XY}^k = \mathcal{P}_X \mathcal{P}_{Y|X}^k$ for $k = 1, \ldots, K$. We use each joint image-label distribution to form an individual MDP environment M_k. The distribution of MDPs $\mathcal{P}_M$ is thus over multiple observers for the same target task. In the multi-observer framework, weights of the task predictor are synced between the different environments at the start of a new trial and the task predictor is updated using Reptile (Nichol et al., 2018) to allow for adaptability and data efficiency of the predictor in addition to the controller.

The task predictor update therefore consists of two steps: 1) perform gradient descent to update the weights of the task predictor $f(\cdot; w_t)$, from w_t to $w_{t,\text{new}}$; 2) update the task predictor weights $w_t \leftarrow w_t + \epsilon(w_{t,\text{new}} - w_t)$. ϵ is set as 1.0 initially and linearly annealed to 0.0 as trials progress.

2.5. Controller selection at inference

A separate holdout set is used to evaluate the controller-learned task amenability assessment. We remove a proportion of the samples valued least by the controller by sorting the samples in the holdout set according to controller predicted values. This ratio of samples removed based on controller outputs is referred to as the 'holdout set rejection ratio'.

3. Experiments

We conduct experiments to evaluate the efficacy of the proposed task amenability framework, both in single and multiple environment settings. The details of the experiment and the data used to evaluate the proposed framework are outlined in this section.

3.1. Experiment data

We used two datasets in our experiments; the first dataset was formed of trans-rectal ultrasound (TRUS) images of the prostate gland and surrounding regions, and the second dataset consisted of chest X-ray images. The TRUS dataset allows us to evaluate the efficacy of the proposed framework to two common surgical guidance and navigational target tasks of prostate presence detection and gland segmentation. The chest X-ray dataset allows us to evaluate the framework for a common diagnostic target task of pneumonia detection using chest X-ray images, and to demonstrate that the proposed approach is not limited to a single imaging modality, dataset, and target task. Moreover, this dataset is publicly available and commonly used for medical imaging research.

3.1.1. Transrectal ultrasound imaging and the target tasks

Ultrasound guided biopsy procedures were performed for 259 patients as part of the clinical trials (NCT02290561, NCT02341677). Trans-rectal ultrasound images were acquired for these patients using a side firing transducer of a bi-plane trans-perineal ultrasound probe (C41L47RP, HI-VISION Preirus, Hitachi Medical Systems Europe). These images were acquired either while manually positioning a digital trans-perineal stepper (D&K Technologies GmbH, Barum, Germany) for navigation using ultrasound or while rotating the stepper with recorded relative angles to scan the entire gland. Each image consisted of 50–120 2D frames of TRUS with relative angles recorded for each frame.

Data pre-processing

In order to feasibly label the acquired data, the resulting TRUS images were sampled at approximately 4 degrees resulting in a total of 6712 2D frames of TRUS from 259 patients. These images were randomly split into train, validation and holdout sets with 4689, 1023, and 1000 images from 178, 43, and 38 subjects, respectively. All the images were labelled by four observers for two downstream target tasks. In addition to the labels for target task, human labels of task amenability for both target tasks were acquired for the purpose of comparing to controller predicted task amenability assessment and in order to select data for the fixed clean validation set reward strategy. These task amenability labels were in the form of binary labels indicating if an image was amenable for the target task or not.

Two target tasks and their labels

Labels for two target tasks were curated for all the images: 1) prostate presence classification (binary scalar value indicating prostate presence); 2) prostate gland segmentation (binary image mask of the prostate gland).

Three sets of labels were collected from three trained medical imaging researchers for both target tasks. For brevity, these three sets are referred to as "non-expert" labels and are denoted by $\{L_i\}_{i=1}^3$. An additional set of consensus labels were generated using a slice-level and pixel-level majority vote for the classification and segmentation tasks, respectively. These consensus labels are denoted by L_C. These consensus labels were then reviewed by a urologist and edited as required. The labels curated by reviewed consensus with an "expert" are denoted as L_R.

Summary of experiments

With this dataset we perform two types of experiments: 1) experiments where we use the consensus labels L_C as a single environment for training using the single environment simplified case presented in Sect. 2.3; 2) experiments where we use the non-expert label sets $\{L_i\}_{i=1}^3$ for training and the expert label set L_R for adaptation using the multi-environment framework (more specifically the multi-observer setting presented in Sect. 2.4). For experiments that fall under type 1, referred to as single-environment experiments, we investigate the effect of the three proposed reward strategies, presented in Sect. 2.2.4, and compare with a non-selective baseline (Sect. 3.4.1), compare controller-predicted task amenability to human labels of IQA (Sect. 3.4.1.2), conduct a sensitivity analysis for the s_{rej} hyper-parameter for the segmentation task (Sect. 3.4.1.1). For experiments that fall under type 2, referred to as multi-environment experiments, we compare the proposed multi-environment meta-RL framework to a single environment baseline (Sect. 3.5), and conduct ablation studies to evaluate the design choices of the proposed meta-RL algorithms and its training strategy (Sect. 3.5).

3.1.2. Public chest X-ray data and the target task

A public dataset of chest X-ray images with binary labels indicating pneumonia diagnosis (Yang et al., 2021) was used to evaluate the applicability of the proposed framework to diagnostic target tasks. A total of 5856 images were randomly split into train, validation and holdout sets with 4708, 524 and 624 images, respectively. The images are scaled down paediatric chest X-ray images acquired as part of routine clinical visits. Since low-quality images were manually removed from this data we added artificial corruptions of random intensities to these images in order to evaluate the task amenability assessment framework. Two corruption operations with random intensities were used. Firstly, random Gaussian noise was added to images with intensities varying between 0.0 and 0.8 where an intensity of 1.0 means all pixels in the image are corrupted by random noise and 0.0 means that no pixels are corrupted (percentage of pixels corrupted in the image varies linearly with intensity value). Secondly, random obstructions were added to the image with intensities between 0.0 and 1.0 where an intensity of 1.0 means 100% of the image is obstructed by zero intensity pixels and 0.0 means that the original image is fully visible with no obstructions added (percentage of image obstructed varies linearly with intensity value).

The dataset used and the associated experiments are available in an open-source GitHub repository: https://github.com/s-sd/task-amenability.

Summary of experiments

With this dataset we perform experiments with the reduced single environment (Sect. 2.3) and thus experiments for this data all fall under single-environment experiments. We evaluate different RL algorithms within the proposed framework (Sect. 3.4.3), and compare the performance of the task predictor when controller selection is applicable to the holdout set with random selection and non-selective baselines (Sect. 3.4.2).

3.2. IQA Performance evaluation

The IQA system, including both the controller and target task predictor are evaluated jointly using task performance. This serves as a direct measure of performance for the task predictor and indirect measure of performance for the controller with respect to the learned task amenability. The task performance measures for the different target tasks considered in this work are outlined as follows.

Prostate presence detection For the prostate presence detection classification task, we report the mean accuracy (Acc.) for the holdout set with standard deviation (St.D.) as a measure of inter-patient variability.

Prostate gland segmentation We report the mean binary Dice score (Dice) on the holdout set with St.D. to measure inter-patient variability, for the prostate gland segmentation task.

Pneumonia detection For the pneumonia detection classification task, we report Acc. and we use bootstrap sampling to estimate the St.D. as a measure of inter-sample variance for the holdout set.

All results, for all tasks, are averaged across all the samples in the holdout set. Where controller selection on the holdout set is applicable, the holdout set rejection ratio is specified and the task performance measures are computed over the selected samples only. Paired _t_-test results at a significance level of 5% are reported for comparisons.

3.3. Training details

Prostate presence detection For the single-environment class of experiments, the controller is trained using the deep deterministic policy gradient (DDPG) (Lillicrap et al., 2019) algorithm with hyper-parameters being empirically configured. For the multi-environment type of experiments the controller is trained using proximal policy optimisation (PPO) (Schulman et al., 2017). An Alex-Net-style architecture (Krizhevsky et al., 2012) was used as the target task predictor. The target task predictor and controller were trained with a cross-entropy loss and reward based on classification accuracy (Acc.) for the classification task. Hyper-parameters for the controller and task predictor network architectures and training procedures remain unchanged from default unless specified. The actor and critic networks used in the DDPG and PPO algorithms pass the image inputs via a 3-layered convolutional encoder. These networks then feed into 3 fully connected layers (which embed an RNN in the case of PPO as detailed in Sect. 3.5).

Prostate gland segmentation The DDPG and PPO algorithms were used for the single- and multi-environment experiments, respectively, similar to the prostate presence detection task and hyper-parameters for these controller networks are also identical between these two tasks. The U-Net architecture (Ronneberger et al., 2015) was used as the task predictor and was trained with a pixel-wise cross-entropy loss. The reward was based on mean binary Dice score.

Pneumonia detection All details including the target task loss, reward and RL hyper-parameters remain the same as the prostate presence classification task, however since there are no multi-environment experiments for this dataset, for comparison between different RL algorithms, both DDPG and PPO are used in the single-environment setting for controller training.

3.4. Single-environment experiments

3.4.1. Evaluating different reward strategies

The proposed RL-based task amenability framework for a single environment was evaluated using the TRUS data with consensus labels L_C for training for all three reward strategies presented in Sec 2.2.4.

To evaluate the relationship of the controller output with the target task performance, different percentages of the holdout set were removed according to controller output and the mean performance measure and measure of spread are reported for the remaining samples. The three reward formulations were compared with a baseline target task predictor with no controller selection during training or testing. For the fixed clean validation set reward formulation, s_{rej} was set as 0.05 and 0.15 for the prostate classification and segmentation tasks, respectively, for all experiments, unless specified otherwise. The holdout set rejection ratio for the prostate presence detection and gland segmentation tasks are set to 0.05 and 0.15, respecitvely.

Sensitivity analysis for the rejection ratio for the selective reward formulation

The selective reward formulation has an additional hyper-parameter s_{rej}. We conducted a sensitivity analysis for this hyper-parameter for the segmentation task for the TRUS data. The performance measure and spread are reported for this task for varying s_{rej} using the selective reward formulation. For the purpose of comparison, the task performance is reported for a holdout set rejection ratio of 0.15 for all tested values of s_{rej}.

Comparing controller labels to human labels of task amenability

We compared human labels of task amenability with controller predictions for the two target tasks for the TRUS data. These comparisons were made for all three reward strategies and for the purpose of comparing with binary human-labelled task amenability, 5% and 15% of the lowest valued controller samples were considered as having low-task amenability. These comparisons are presented in the form of contingency tables.

3.4.2. Controller selection and a non-selective baseline

The weighted validation set strategy was used to evaluate the efficacy of the proposed RL-based task amenability framework on the chest X-ray dataset, without using any human labels of task amenability. The relationship between target task performance and controller output was studied by removing a proportion of samples based on controller output. The results are reported for a holdout set rejection ratio of 0.10. Two baseline networks trained and tested, 1) without controller selection referred to as the non-selective baseline, and 2) with a random selection strategy, i.e. a trivial controller, referred to as the random selection baseline, are quantitatively compared.

3.4.3. Comparing different RL algorithms

To evaluate the sensitivity of the proposed framework to different RL algorithms. The proximal policy optimisation (PPO) and deep deterministic policy gradient (DDPG) algorithms were compared for controller training, based on the weighted validation set reward strategy with the chest X-ray dataset. The holdout set rejection ratio for this experiment was set as 0.10.

3.5. Multi-environment for multi-observer labelling

To evaluate the multi-environment meta-RL framework, for the multi-observer setting, we trained and compared three models for task amenability:

- _Meta-baseline_: This model was trained with all the high-quality reviewed consensus labels L_R. Only a single environment was used to train with this "expert" labelled data to establish a reference.
- _Meta-RL_: The proposed meta-RL framework was used for training with the three non-expert labels $\{L_i\}_{i=1}^3$ forming the three environments for training where the train and validation sets are used. The task predictor and controller were subsequently adapted using $k \times 100\%$ of the training and validation sets with expert labels L_R.
- _Meta-RL Variant_: We conduct an ablation study in order to evaluate the effectiveness of the environment-level separation, we trained a model using all of the training and validation data using non-expert labels $\{L_i\}_{i=1}^3$ as a single environment. That is to say that the trials were not environment specific and Reptile was not used to perform updates to optimise the task predictor as it reduces to gradient descent in this case. The task predictor and controller were adapted using $k \times 100\%$ of the training and validation sets with expert labels L_R but for this model, the RNN internal state was not reset before adaptation.

Table 1
Results on the controller-selected holdout set.

Task	Reward computation strategy	Mean ± St.D.
Prostate presence	Non-selective baseline	0.897 ± 0.010
(Acc.)	$\tilde{R}_{avg,t}$, fixed validation set	0.935 ± 0.014
	$\tilde{R}_{w,t}$, weighted validation set	0.926 ± 0.012
	$\tilde{R}_{sel,t}$, selective validation set	0.913 ± 0.012
Prostate segmentation	Non-selective baseline	0.815 ± 0.018
(Dice)	$\tilde{R}_{avg,t}$, fixed validation set	0.890 ± 0.017
	$\tilde{R}_{w,t}$, weighted validation set	0.893 ± 0.018
	$\tilde{R}_{sel,t}$, selective validation set	0.865 ± 0.014
Pneumonia detection	Non-selective baseline	0.817 ± 0.026
(Acc.)		
	$\tilde{R}_{w,t}$, weighted validation set (DDPG)	0.843 ± 0.033
	$\tilde{R}_{w,t}$, weighted validation set (PPO)	0.838 ± 0.031

We evaluate these models for varying k-values, where k is the ratio of expert-labelled samples used for adaptation ($k \times 100\%$ samples used).

The TRUS data was used for this experiment and the multi-environment framework was evaluated for both target tasks. The controller, for this experiment, was embedded with an RNN and had additional inputs of the previous reward, action and terminal flag. The image was passed through 3 convolutional layers before being passed to the RNN embedded controller which had a stacked LSTM architecture with hyper-parameters remaining unchanged from defaults in (Wang et al., 2017). PPO was used to train the controller in this meta-RL framework since the DDPG algorithm relies on random sampling from a replay buffer which is not well suited when adaptation is done based on RNN internal state updates, which builds sequential memory in the system.

4. Results

Reward strategies The results of Acc. and Dice from the TRUS dataset, with the three proposed reward strategies, are summarised in Table 1. The performance measures are computed after a controller selection of the holdout set with holdout set rejection ratios of 0.05 and 0.15 for the prostate presence classification and prostate segmentation tasks, respectively. For both of the tasks, all three proposed reward formulations were able to achieve higher performance compared to a non-selective baseline, with statistical significance (*p-values<0.001*). The fixed clean validation set and weighted validation set strategies led to significantly higher performance (*p-values<0.001*) compared to the selective validation set reward formulation, for both tested tasks. However, significance was not observed between those from the fixed clean validation set and weighted validation set reward strategies, for both the classification (*p-value=0.06*) and segmentation (*p-value=0.49*) tasks. To evaluate performance at varying holdout set rejection ratios, the mean performance against holdout set rejection ratio is plotted in Fig. 3. The peak classification Acc. 0.935, 0.932 and 0.913 occur at 5%, 10% and 5% rejection ratios, for the fixed-, weighted- and selective reward formulations, respectively, while the peak segmentation Dice 0.891, 0.893 and 0.866 occur at 20%, 15% and 20% rejection ratios, respectively.

Comparison to human IQA labels Contingency tables are presented in Fig. 4 to compare controller predicted task amenability to human labels of task-specific quality for the holdout set, for the prostate presence classification and prostate segmentation tasks. For the purpose of comparison, the holdout set rejection ratio is set to 0.05 and 0.15, for the classification and segmentation tasks, respectively, such that rejected samples are considered low task amenability and the rest are considered high task amenability. Agreement in low task amenability samples of 75%, 70% and 43%, with Cohen's kappa values of 0.75, 0.51, was seen for the prostate

Table 2
Sensitivity analysis for s_{rej} for the prostate segmentation task for the selective reward formulation.

s_{rej}	Mean ± St.D.
0.00	0.827 ± 0.011
0.05	0.838 ± 0.012
0.10	0.845 ± 0.010
0.15	0.865 ± 0.014
0.20	0.882 ± 0.017
0.25	0.888 ± 0.016
0.30	0.876 ± 0.012

presence classification task, for the fixed-, weighted- and selective validation sets, respectively. In the segmentation task, agreement on low task amenability samples of 65%, 58% and 49%, with Cohen's kappa values of 0.63, 0.48 and 0.37, was observed for the three reward formulations, respectively.

Sensitivity analysis and ablation studies The validation set rejection ratio s_{rej} is treated as a hyper-parameter and a sensitivity analysis for this hyper-parameter is conducted for the prostate segmentation task. The performance at varying values of s_{rej} is presented in Table 2. s_{rej} is increased in increments of 0.05 and each step increase leads to a statistically significant improvement in performance up to $s_{rej} = 0.20$ (*p-values<0.01*). No significance was found comparing performances for a step increase of s_{rej} from 0.20 to 0.25 (*p-value=0.37*). A subsequent step increase from 0.25 to 0.3 led to a statistically significant performance reduction (*p-value<0.01*).

To evaluate the sensitivity of the proposed framework to different RL algorithms, we present task performance results for the pneumonia detection task for a holdout set rejection ratio of 0.10 in Table 1. Although DDPG showed improved performance, significance was not observed (*p-value=0.20*) for the difference between the two algorithms. Both the DDPG- and PPO-trained controller showed significantly improved performance compared with the non-selective baseline (*p-values<0.001*). To evaluate the relationship between performance and holdout set rejection ratio, we present a plot of mean performance against holdout set rejection ratio in Fig. 6, which presents a comparison between the task amenability framework, a non-selective baseline and a random selection baseline. An implementation of the proposed framework, with both the PPO and DDPG algorithms, along with the data, code, and results for the pneumonia detection task are available in the same open-source GitHub repository.

Adaptability performance of meta-RL The proposed meta-RL multi-environment framework was evaluated using the prostate classification and segmentation tasks and the results are summarised in Table 3. Additionally, a plot of performance against varying k values is presented in Fig. 7 for both tasks. Statistical significance was not found between the baseline and meta-RL for
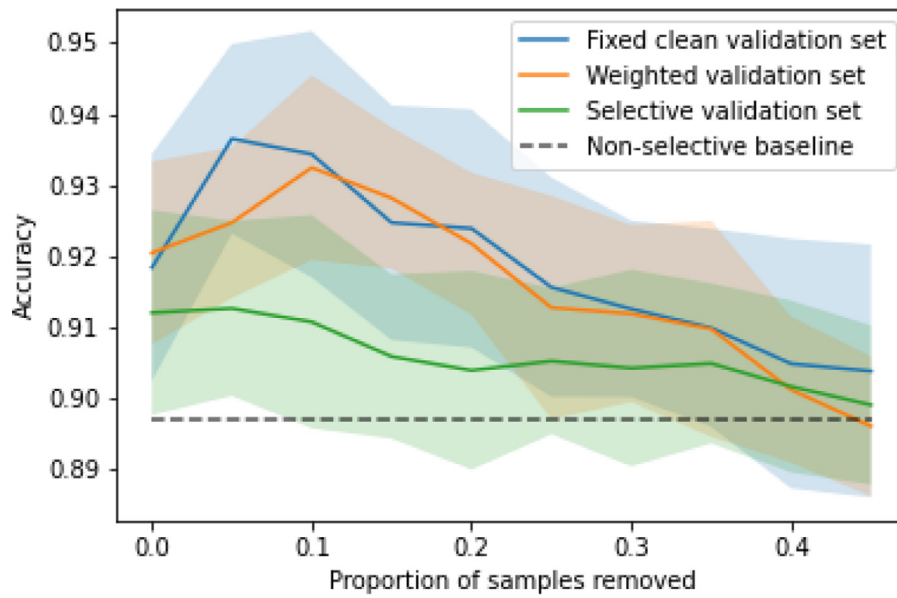
(a) Prostate presence classification task
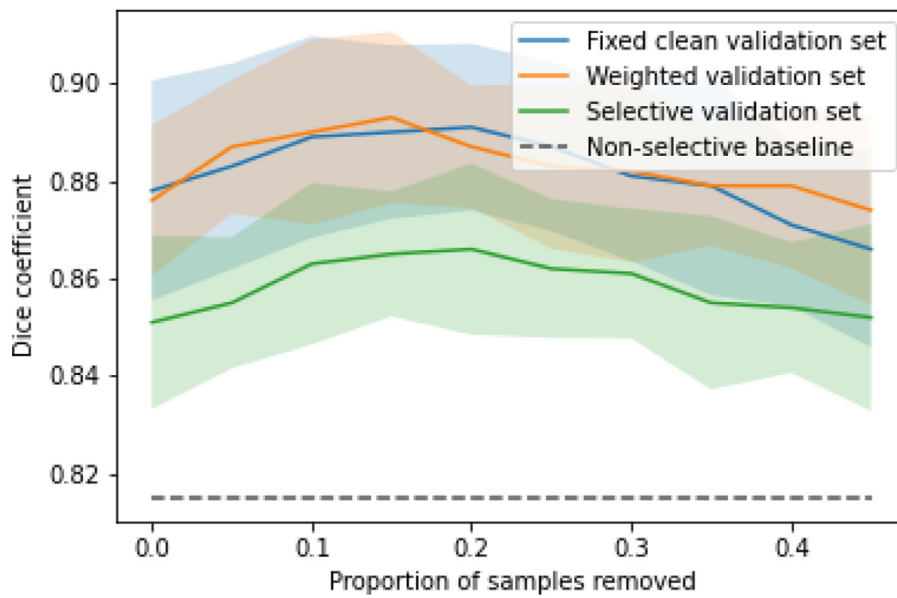
(b) Prostate segmentation task

Fig. 3. Plots of the task performance (in respective Acc. and Dice metrics) against the holdout set rejection ratio for the two tasks. (Colour figure).

Table 3
Comparison of holdout set results with a rejection ratio set to 0.05 (Meta-RL).

Tasks		Prostate Classification (Acc.)	Prostate Segmentation (Dice)
IQA Methods	k	Mean ± St.D.	Mean ± St.D.
Meta-baseline	N/A	0.932 ± 0.011	0.894 ± 0.016
Meta-RL	0.5	0.936 ± 0.012	0.892 ± 0.018
	0.4	0.929 ± 0.016	0.886 ± 0.014
	0.3	0.926 ± 0.010	0.888 ± 0.020
	0.2	0.925 ± 0.017	0.873 ± 0.017
	0.1	0.911 ± 0.012	0.863 ± 0.020
	0.0	0.908 ± 0.010	0.857 ± 0.018
Meta-RL	0.5	0.931 ± 0.015	0.884 ± 0.016
Variant	0.4	0.920 ± 0.010	0.882 ± 0.021
	0.3	0.919 ± 0.013	0.882 ± 0.015
	0.2	0.916 ± 0.014	0.860 ± 0.014
	0.1	0.905 ± 0.014	0.858 ± 0.021
	0.0	0.896 ± 0.016	0.849 ± 0.017

Fig. 4. Contingency tables comparing subjective labels to controller predictions for the different reward computation strategies. (Colour figure).
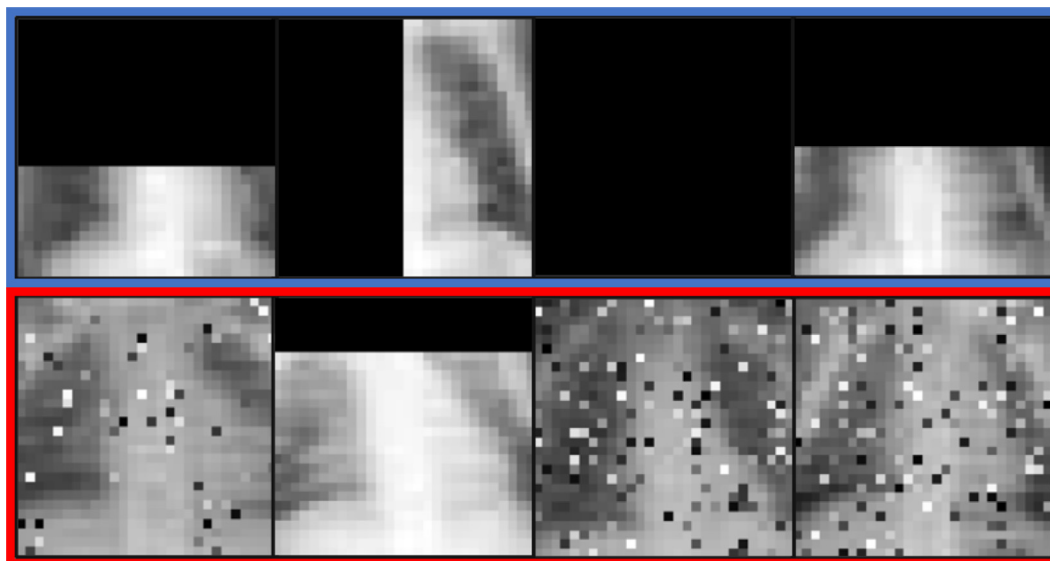
Fig. 5. Controller prediction samples for the pneumonia detection task. **Blue**: samples predicted as having low task amenability; **Red**: samples predicted as having high task amenability. Low task amenability refers to controller predicted value below the 10-th percentile and high refers to above. More examples can be found with the open-source repository. (Colour figure).
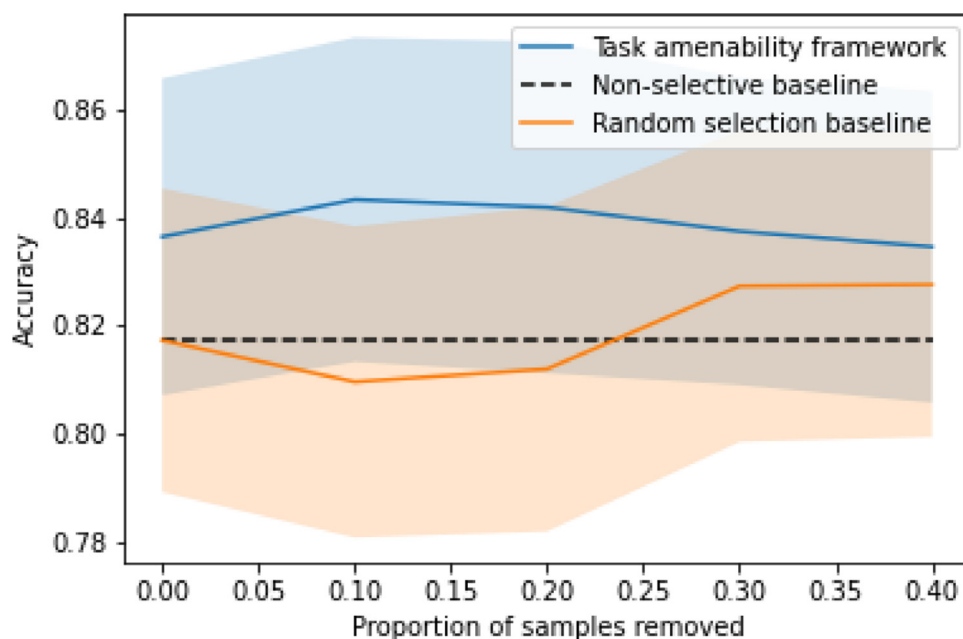
Fig. 6. Plot of performance (Acc.) against holdout set rejection ratio for pneumonia detection task. (Colour figure).
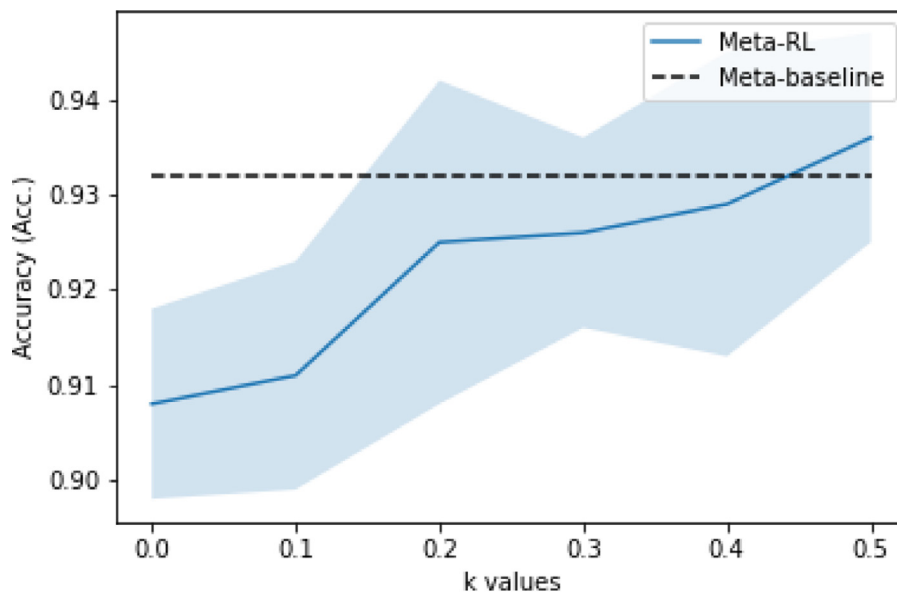
(a) Prostate presence classification task
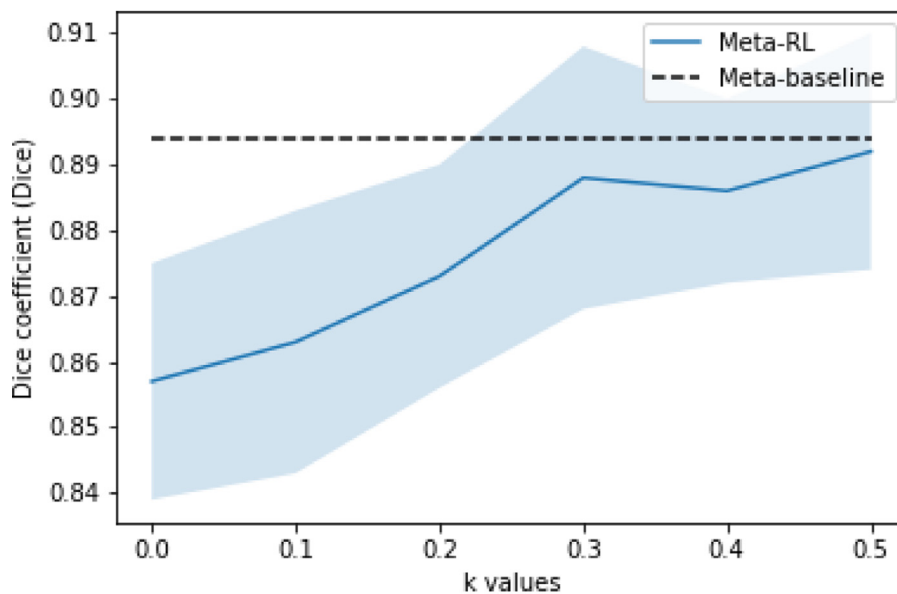
(b) Prostate segmentation task

Fig. 7. Plots of the task performance (in respective Acc. and Dice metrics) against the *k* values with a rejection ratio set to 5%. (Colour figure).

the prostate presence classification task, with *k* values from 0.5 to 0.2 (*0.10<p-values<0.23*), however, the baseline showed significantly higher performance compared to meta-RL for low k-values of 0.1 and 0.0 (*p-values<0.01* for both). For the prostate segmentation task, no statistical significance was found between the baseline and meta-RL for *k*-values from 0.5 to 0.3 (*0.07<p-values<0.17*) but the baseline performance was significantly higher than meta-RL for low *k* values from 0.2 to 0.0 (*p-values<0.01*). For the ablation study, the proposed meta-RL framework outperformed the meta-RL variant which had no environment-level separation, with statistical significance, for *k* values from 0.0 to 0.4 (*p-values<0.01*), for the classification task and for all *k* values, for the segmentation task (*p-values<0.03*). A significant difference was not observed for a high *k* value of 0.5 in the classification task (*p-value*=0.06). In an-

other ablation study, compared to a Reptile-omitted meta-RL variant, which achieved Acc.=0.901 ± 0.013 and Dice=0.851 ± 0.013 for the classification and segmentation tasks, respectively, the meta-RL framework showed higher performance, with statistical significance (*p-values<0.01*), for both tasks for a *k*-value of 0.0. For all other *k*-values no significant differences were seen between meta-RL and the its Reptile-omitted variant.

5. Discussion

The results presented in Sect. 4 show that the proposed task amenability framework is able to offer an increased performance, compared to a non-selective baseline for all tested tasks which demonstrates the efficacy of the proposed approach. The three
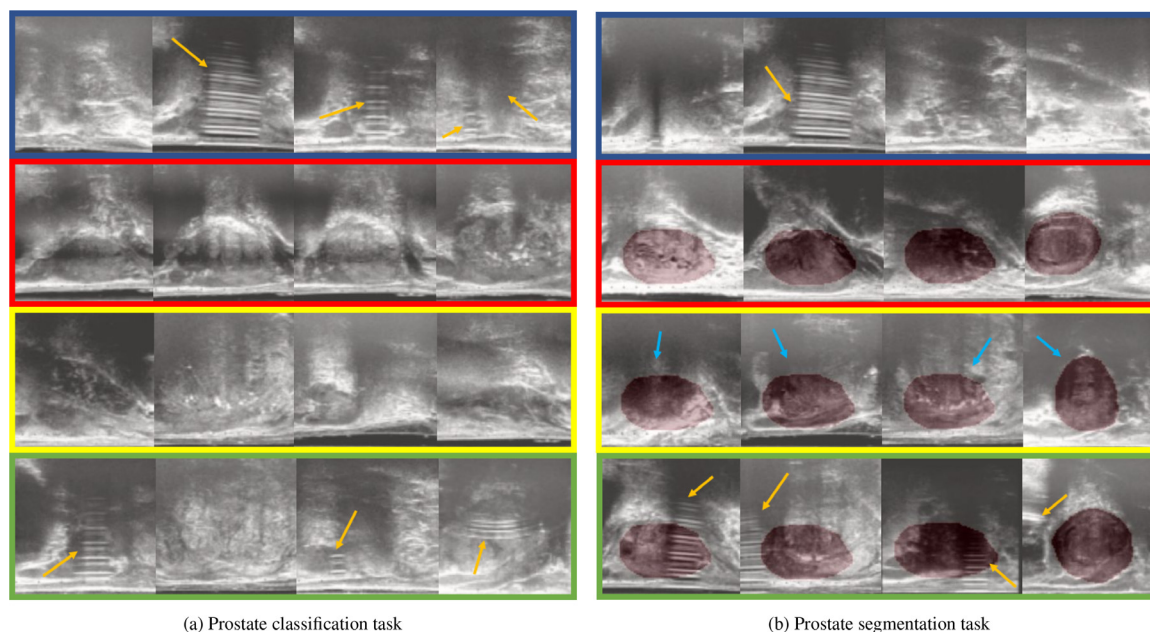
(a) Prostate classification task   (b) Prostate segmentation task

Fig. 8. Examples of controller selected and rejected images (rejection ratio=5%) for both tasks for the multi-environment framework. **Blue:** rejected samples; **Red:** selected samples; **Yellow:** rejected samples despite no apparent artefacts or severe noise; **Green:** selected samples despite present artefacts or low contrast. **Orange arrows:** visible artefacts; **Cyan arrows:** regions where gland boundary delineation may be challenging. (Colour figure).

tasks include a range of clinical applications, including prostate presence classification, prostate gland segmentation and pneumonia detection, for surgical navigation, interventional guidance and diagnostic assistance.

It was found particularly useful that task amenability can be learnt within the proposed IQA framework without any human labels of image quality, by using the weighted and selective reward formulations. While only the weighted reward formulation offers comparable performance to the fixed clean validation set reward formulation, which requires human labels of image quality, the selective formulation provides a means to specify a clinically desirable rejection rate. Furthermore, tuning this validation set rejection ratio parameter, s_{rej}, for the selective formulation may be necessary in order to achieve performance comparable to the weighted- and fixed clean validation set reward formulations. In practice, we observed that increasing s_{rej} beyond a certain point, the performance was significantly reduced which could potentially be due to over-fitting of the agent, caused by an increased exploration space as a result of the increased number of samples to be rejected from the validation set. This parameter could thus also impact the exploration-exploitation trade-off in training the proposed task amenability framework. Interestingly, the mean performance showed a small decrease after an initial rise, with increasing hold-out set rejection ratio, for all tested reward formulations. The variance of predictions, possible over-fitting of the RL agent and the overall high quality of the datasets used (which could in turn limit the overall performance improvement), may offer some explanation for this observation, however, it still remains an open question.

The multi-environment framework, proposed to learn adaptable task amenability assessment, is potentially applicable to several scenarios, for example, learning an adaptable IQA definition over different target tasks or task predictor functions, besides different datasets. This multi-environment framework was investigated in the clinically relevant scenarios to equip adaptability, over different observer labels, to both the controller and task predictor functions. Adaptation to an expert labelling standard was achieved with 20–30% of the expert labelled data without any significant reduc-

tion in performance compared to using 100% of the expert labelled data for training. This meant that, for the classification and segmentation tasks, 1087 and 1634 expert-labelled images from 42 and 63 subjects (training and validation sets), respectively, were sufficient to achieve performance comparable to using 100% of the expert-labelled data for training. It is important to note that the multi-environment formulation also required non-expert-labelled data for training, however, these labels may be used to learn varying definitions of task amenability; further economic analysis of the use of non-expert data is beyond the scope of this work. The applicability of the framework, to equip adaptability to allow for efficient use of expert-labelled data, brings to light other potential use cases such as to equip adaptability across different target tasks to adapt IQA definitions to new tasks using limited labelled data which could be useful for several applications such as active learning, and particularly of interest in developing resource-constrained clinical applications.

6. Conclusion

This work introduces task amenability as an alternative to traditional subjective definition of image quality, especially for downstream machine learning tasks. We also propose a mechanism to efficiently adapt such RL-based IQA agents to new labelling standards. Learning a task-specific IQA or task amenability is useful for several applications such as for re-acquisition guidance for applications such as ultrasound where re-acquisition is inexpensive, for operator skill feedback and for meeting clinically defined accuracy requirements for downstream clinical tasks. The proposed multi- and single-environment frameworks promise a new method to learn such a task-specific IQA or task amenability. The proposed approach has been shown its wide applicability with clinically relevant target tasks of prostate presence detection and gland segmentation, and pneumonia detection, with real clinical data from prostate cancer and pneumonia patients, respectively.

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.

Acknowledgements

References

Abdi, A.H., Luong, C., Tsang, T., Allan, G., Nouranian, S., Jue, J., Hawley, D., Fleming, S., Gin, K., Swift, J., Rohling, R., Abolmaesumi, P., 2017. Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. IEEE Trans. Med. Imag. 36 (6), 1221–1230. doi:10.1109/TMI.2017.2690836.

Baum, Z., Bonmati, E., Cristoni, L., Walden, A., Prados, F., Kanber, B., Barratt, D., Hawkes, D., Parker, G., Wheeler-Kingshott, C., Hu, Y., 2021. Image quality assessment for closed-loop computer-assisted lung ultrasound. In: Linte, C.A., Siewerdsen, J.H. (Eds.), Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. International Society for Optics and Photonics. SPIE, pp. 160–166. doi:10.1117/12.2581865.

Botvinick, M., Ritter, S., Wang, J., Kurth-Nelson, Z., Blundell, C., Hassabis, D., 2019. Reinforcement learning, fast and slow. Trends Cogn. Sci. (Regul. Ed.) 23 (5), 408–422.

Brown, A.M., Elbuluk, O., Mertan, F., Sankineni, S., Margolis, D.J., Wood, B.J., Pinto, P.A., Choyke, P.L., Turkbey, B., 2015. Recent advances in image-guided targeted prostate biopsy. Abdom. Imag. 40 (6), 1788–1799.

Camps, S., Houben, T., Carneiro, G., Edwards, C., Antico, M., Dunnhofer, M., Martens, E., Baeza, J., Vanneste, B., van Limbergen, E., et al., 2020. Automatic quality assessment of transperineal ultrasound images of the male pelvic region, using deep learning. Ultrasound Med. Biol. 46 (2), 445–454.

Choong, M.K., Logeswaran, R., Bister, M., 2006. Improving diagnostic quality of mr images through controlled lossy compression using spiht. J. Med. Syst. 30 (3), 139–143.

Chow, L., Paramesran, R., 2016. Review of medical image quality assessment. Biomed. Signal Process. Control 27, 145–154.

Chow, L., Rajagopal, H., 2015. Comparison of difference mean opinion score (dmos) of magnetic resonance images with full-reference image quality assessment (fr-iqa). Image Processing, Image Analysis and Real-Time Imaging (IPARTI) Symposium.

Cubuk, E., Zoph, B., Mane, D., Vasudevan, V., Le, Q., 2019. Autoaugment: Learning augmentation policies from data. 1805.09501.

Daly, S.J., 1992. Visible differences predictor: an algorithm for the assessment of image fidelity. In: Human Vision, Visual Processing, and Digital Display III. Vol. 1666. International Society for Optics and Photonics, pp. 2–15.

Davis, H., Russell, S., Barriga, E., Abramoff, M., Soliz, P., 2009. Vision-based, real-time retinal image quality assessment. In: 2009 22nd IEEE Int. Symp. on Computer-Based Medical Systems, pp. 1–6.

De Angelis, A., Moschitta, A., Russo, F., Carbone, P., 2007. Image quality assessment: an overview and some metrological considerations. In: 2007 IEEE International Workshop on Advanced Methods for Uncertainty Estimation in Measurement. IEEE, pp. 47–52.

Dietrich, O., Raya, J.G., Reeder, S.B., Reiser, M.F., Schoenberg, S.O., 2007. Measurement of signal-to-noise ratios in mr images: influence of multichannel coils, parallel imaging, and reconstruction filters. J. Magnetic Resonance Imag. 26 (2), 375–385.

Dirix, P., Haustermans, K., Vandecaveye, V., 2014. The value of magnetic resonance imaging for radiotherapy planning. In: Seminars in radiation oncology, Vol. 24. Elsevier, pp. 151–159.

Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imag. Graph. 31 4–5, 198–211.

Duan, Y., Schulman, J., Chen, X., Bartlett, P., Sutskever, I., Abbeel, P., 2016. Rl^2: Fast reinforcement learning via slow reinforcement learning. 1611.02779.

Dueholm, M., 2006. Transvaginal ultrasound for diagnosis of adenomyosis: a review. Best Pract. Res. Clin. Obstetric. Gynaecol. 20 (4), 569–582. doi:10.1016/j.bpobgyn.2006.01.005. Adenomyosis

Dutta, J., Ahn, S., Li, Q., 2013. Quantitative statistical methods for image quality assessment. Theranostics 3 (10), 741.

Eck, B.L., Fahmi, R., Brown, K.M., Zabic, S., Raihani, N., Miao, J., Wilson, D.L., 2015. Computational and human observer image quality evaluation of low dose, knowledge-based ct iterative reconstruction. Med. Phys. 42 (10), 6098–6111.

Esses, S., Lu, X., Zhao, T., Shanbhogue, K., Dane, B., Bruno, M., Chandarana, H., 2018. Automated image quality evaluation of t2-weighted liver mri utilizing deep learning architecture. J. Magn. Reson. Imag. 47 (3), 723–728.

Fuderer, M., 1988. The information content of mr images. IEEE Trans. Med. Imag. 7 (4), 368–380.

Geissler, A., Gartus, A., Foki, T., Tahamtan, A.R., Beisteiner, R., Barth, M., 2007. Contrast-to-noise ratio (cnr) as a quality parameter in fmri. J. Magnetic Resonance Imag. 25 (6), 1263–1270.

Hemmsen, M.C., Petersen, M.M., Nikolov, S.I., Nielsen, M.B., Jensen, J.A., 2010. Ultrasound image quality assessment: A framework for evaluation of clinical image quality. In: Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy, Vol. 7629. International Society for Optics and Photonics, p. 76290C.

Henkelman, R.M., 1985. Measurement of signal intensities in the presence of noise in mr images. Med. Phys. 12 (2), 232–233.

Huo, D., Xu, D., Liang, Z.-P., Wilson, D., 2006. Application of perceptual difference model on regularization techniques of parallel mr imaging. Magn. Reson. Imag. 24 (2), 123–132.

Jiang, Y., Huo, D., Wilson, D.L., 2007. Methods for quantitative image quality evaluation of mri parallel reconstructions: detection and perceptual difference model. Magn. Reson. Imag. 25 (5), 712–721.

Kalayeh, M.M., Marin, T., Brankov, J.G., 2013. Generalization evaluation of machine learning numerical observers for image quality assessment. IEEE Trans. Nucl. Sci. 60 (3), 1609–1618.

Kaufman, L., Kramer, D.M., Crooks, L.E., Ortendahl, D.A., 1989. Measuring signal-to-noise ratios in mr imaging. Radiology 173 (1), 265–267.

Kondziolka, D., Lunsford, L.D., 1996. Intraoperative navigation during resection of brain metastases. Neurosurg. Clin. N. Am 7 (2), 267–277.

Kowalik-Urbaniak, I., Brunet, D., Wang, J., Koff, D., Smolarski-Koff, N., Vrscay, E.R., Wallace, B., Wang, Z., 2014. The quest for'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images 9037, 903717.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. NeurIPS.

Kumar, B., Sinha, G., Thakur, K., 2011. Quality assessment of compressed mr medical images using general regression neural network. Int. J. Pure Appl. Sci. Technol. 7 (2).

Kumar, R., Rattan, M., 2012. Analysis of various quality metrics for medical image processing. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 2 (11), 137–144.

Khler, T., Budai, A., Kraus, M.F., Odstrilik, J., Michelson, G., Hornegger, J., 2013. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proc. of the 26th IEEE Int. Symp. on Computer-Based Medical Systems, pp. 95–100.

Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2019. On modelling label uncertainty in deep neural networks: automatic estimation of intra-observer variability in 2d echocardiography quality assessment. IEEE Trans. Med. Imag. 39 (6), 1868–1883.

Lillicrap, T., Hunt, J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2019. Continuous control with deep reinforcement learning. 1509.02971.

Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., Lei, B., 2019. Multi-task learning for quality assessment of fetal head ultrasound images. Med. Image. Anal. 58, 101548. doi:10.1016/j.media.2019.101548.

Liney, G.P., Moerland, M.A., 2014. Magnetic resonance imaging acquisition techniques for radiotherapy planning. In: Seminars in radiation oncology. Vol. 24. Elsevier, pp. 160–168.

Loizou, C.P., Pattichis, C.S., Pantziaris, M., Tyllis, T., Nicolaides, A., 2006. Quality evaluation of ultrasound imaging in the carotid artery based on normalization and speckle reduction filtering. Med. Biol. Eng. Comput. 44 (5), 414–426.

Miao, J., Huo, D., Wilson, D.L., 2008. Quantitative image quality evaluation of mr images using perceptual difference models. Med. Phys. 35 (6Part1), 2541–2553.

Mortamet, B., Bernstein, M.A., Jack Jr, C.R., Gunter, J.L., Ward, C., Britson, P.J., Meuli, R., Thiran, J.-P., Krueger, G., 2009. Automatic quality assessment in structural brain magnetic resonance imaging. Magnetic Resonan. Med. 62 (2), 365–372.

Nichol, A., Achiam, J., Schulman, J., 2018. On first-order meta-learning algorithms. 1803.02999.

Racine, D., Ba, A.H., Ott, J.G., Bochud, F.O., Verdun, F.R., 2016. Objective assessment of low contrast detectability in computed tomography with channelized hotelling observer. Phys. Med. 32 (1), 76–83.

Rangaraju, D.K.S., Kumar, K., Renumadhavi, C., 2012. Review paper on quantitative image quality assessment–medical ultrasound images. Int. J. Eng. Res. Technol. 1 (4).

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. MICCAI. Vol. 9351. Springer.

Saeed, S.U., Fu, Y., Baum, Z.M.C., Yang, Q., Rusu, M., Fan, R.E., Sonn, G.A., Barratt, D.C., Hu, Y., 2021. Learning image quality assessment by reinforcing task amenable data selection. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (Eds.), Infor-

mation Processing in Medical Imaging. Springer International Publishing, Cham, pp. 755–766.

Saeed, S. U., Fu, Y., Stavrinides, V., Baum, Z. M. C., Yang, Q., Rusu, M., Fan, R. E., Sonn, G. A., Noble, J. A., Barratt, D. C., Hu, Y., 2021b. Adaptable image quality assessment using meta-reinforcement learning of task amenability. 2108.04359.

Salem, K.A., Lewin, J.S., Aschoff, A.J., Duerk, J.L., Wilson, D.L., 2002. Validation of a human vision model for image quality evaluation of fast interventional magnetic resonance imaging. J. Electron. Imag. 11 (2), 224–235.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. 1707.06347.

Shiao, Y.-H., Chen, T.-J., Chuang, K.-S., Lin, C.-H., Chuang, C.-C., 2007. Quality of compressed medical images. J. Digit. Imag. 20 (2), 149.

Shima, Y., Suwa, A., Gomi, Y., Nogawa, H., Nagata, H., Tanaka, H., 2007. Qualitative and quantitative assessment of video transmitted by dvts (digital video transport system) in surgical telemedicine. J. Telemed. Telecare 13 (3), 148–153.

Simpson, A., Kingham, T., 2016. Current evidence in image-guided liver surgery. J. Gastrointest. Surg. 20, 1265–1269.

Sinha, A., Malo, P., Deb, K., 2018. A review on bilevel optimization: from classical to evolutionary approaches and applications. IEEE Trans. Evol. Comput. 22, 276–295.

Wang, J., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J., Munos, R., Blundell, C., Kumaran, D., Botvinick, M., 2017. Learning to reinforcement learn. 1611.05763.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Woodard, J.P., Carley-Spencer, M.P., 2006. No-reference image quality metrics for structural mri. Neuroinformatics 4 (3), 243–262.

Wu, L., Cheng, J., Li, S., Lei, B., Wang, T., Ni, D., 2017. Fuiqa: fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybernetics 47 (5), 1336–1349.

Yang, J., Shi, R., Ni, B., 2021. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. 2010.14925.

Yoon, J., Arik, S., Pfister, T., 2020. Data valuation using reinforcement learning. 1909.11671.

Zago, G., Andreo, R., Dorizzi, B., Ottoni, E., Salles, T., 2018. Retinal image quality assessment using deep learning. Comput. Biol. Med. 103, 64–70.

Zhang, X., Wang, Q., Zhang, J., Zhong, Z., 2019. Adversarial autoaugment. 1912.11188.

Zoph, B., Le, Q., 2017. Neural architecture search with reinforcement learning. 1611.01578.