Reinforcing Medical Image Detectors: Concepts, Applications, Challenges and Future Directions

Mingzhe Hu*
mingzhe.hu@emory.edu
Computer and Informatics Department, Emory University
Deep Biomedical Imaging Lab, School of Medicine, Emory University
Atlanta, Georgia, USA

Xiaofeng Yang
xiaofeng.yang@emory.edu
Winship Cancer Institute, Emory University
Radiology Department, School of Medicine, Emory University
Atlanta, Georgia, USA

Figure 1: Winship Cancer Institute, Emory University

ABSTRACT

Motivation: Medical image analysis involves a series of tasks to assist physicians in qualitative and quantitative analysis of lesions or anatomical structures, significantly improving the accuracy and reliability of medical diagnosis and prognosis. Traditionally, these tedious tasks are finished by experienced physicians or medical physicists and lead to two major problems: (i) low efficiency; (ii)biased by personal experience.

In the past decade, many machine learning methods have been applied to accelerate and automate the image analysis process. Compared to the enormous deployments of supervised and unsupervised learning models, the attempts to use reinforcement learning in anatomical landmark detection are still scarce. We hope that this review article could serve as the stepping stone for related research in the future.

Approach & Results: We selected published articles from Google Scholar and PubMed. Considering the scarcity of related articles, we also included some outstanding newest preprints. The papers are carefully reviewed and categorized according to the type of image analysis task. In this article, we first review the basic concepts and popular models of reinforcement learning. Then we explore the applications of reinforcement learning models in various medical image analysis tasks. Finally, we conclude the article by discussing the reviewed reinforcement learning approaches' limitations and possible future improvements.

Significance: From our observation, though reinforcement learning has gradually become a hot topic in recent years, many researchers in the medical analysis field still find it hard to understand and deploy in clinical settings. One cause is lacking well-organized review articles targeting readers lacking professional computer science background. Rather than providing a comprehensive list of all reinforcement learning models applied in anatomical landmark detection, We want to help the readers to learn how to formulate and solve their medical image analysis research as reinforcement learning problems.

*If you need access to the full article, please contact the author.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**;
• **Applied computing** → **Health informatics**.

KEYWORDS

Reinforcement Learning, Medical Image, Survey, Modalities

ACM Reference Format:

Mingzhe Hu and Xiaofeng Yang. 2022. Reinforcing Medical Image Detectors: Concepts, Applications, Challenges and Future Directions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The purpose of medical image analysis is to mine and analyze valuable information from medical images by using digital image processing techniques to assist doctors in making more accurate and reliable diagnoses and prognoses. According to different imaging principles, common imaging modalities can be categorized as CT, MR, Ultrasound, SPECT, PET, X-ray, OCT, and microscope. Medical image processing can also be classified according to specific processing tasks. Typical tasks include classification, segmentation, registration, and recognition. To present readers with a comprehensive review, we extend the image analysis concept a little bit to include image reconstruction, image synthesis, and radiotherapy planning. Figure 2 shows the range of our review article.

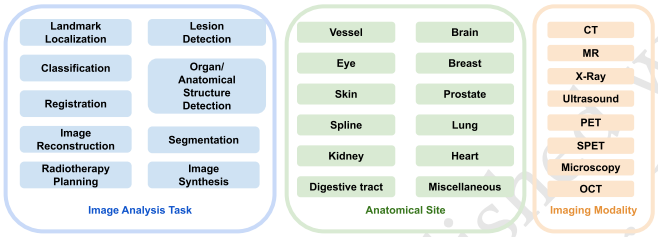

Figure 2: Range of our review article. Blue box: covered image analysis tasks; green box: covered anatomical sites; yellow box: covered imaging modalities.

With the development of imaging technology and the iterative update of imaging equipment, the time required for medical imaging is greatly shortened, and the resolution of imaging is also significantly improved. At the same time, the data volume of medical images has experienced an unprecedented surge, with the trend of high-dimensionality. The traditional manual analysis of medical images by physicians became tedious and inefficient. More and more physicians are looking to automate this process by partnering with engineers and data scientists. That's how the combined medical imaging and machine learning field was born. Many excellent algorithms in the field of natural image analysis have also shown good results in the field of medical images [Shen et al. 2017].

Reinforcement learning (RL) is neither supervised learning nor unsupervised learning. The goal of reinforcement learning is to achieve the maximum expected cumulative reward [Sutton and Barto 2018]. Reinforcement learning Figure 3 shows the relationship between machine learning, supervised learning, unsupervised learning, and deep learning.

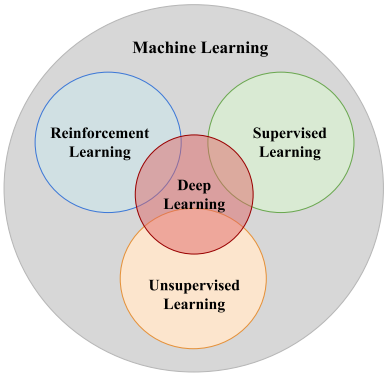

Figure 3: Relationship between machine learning, supervised learning, unsupervised learning and reinforcement learning.

The number of published reinforcement learning-related papers has grown rapidly in the past 2 decades. State-of-the-art RF models have been applied to solve problems in the game [Vinyals et al. 2019], natural language processing [Sharma and Kaushik 2017], autonomous driving [Sallab et al. 2017] that are hard to be solved using other machine learning approaches and achieved outstanding performance. However, attempts to exploit RF in the medical analysis field are still scarce compared to these applications. Figure 4 shows the trends of number of published machine learning papers and reinforcement learning papers in medical image analysis. Despite the overall growth trend, the number of published reinforcement learning papers still only constitutes a tiny part of machine learning in medical image analysis. On the other hand, RF

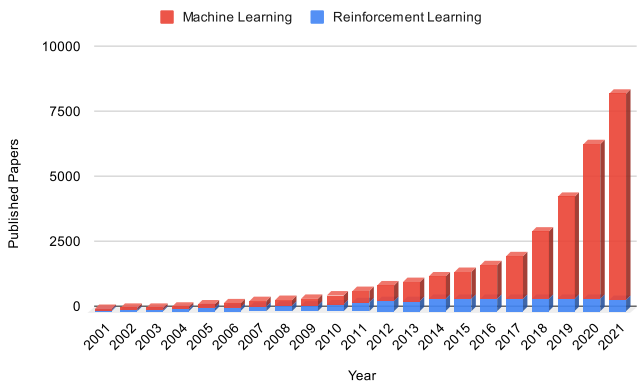

Figure 4: Trends of number of published machine learning papers and reinforcement learning papers in medical image analysis. Data Source: PubMed.

methods have unique advantages dealing with medical image data:

- Training the RF models does not require lots of labeled data, while medical image data often lacks large-scale accessible annotation.

- RF models are less biased since they won't Inherit bias from the labels made by human annotators.
- RF can learn from sequential data, and the learning process is goal-oriented. Besides exploiting experience, it can also explore new solutions. The RF can even surpass human experts when solving the same problem.

The review article is be based on Synthesis Methodology [Wilson and Anagnostopoulos 2021]. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) will be followed [Moher et al. 2009]. Firstly, the following pattern will be searched in Google Scholar and PubMed: (Reinforcement Learning) AND [Imaging Modality] AND [Analysis Task]. Here [Imaging Modality] = (CT OR MR OR OCT OR Ultrasound OR X-ray OR SPECT OR PET), and [Analysis Task] = (Segmentation OR Classification OR Registration OR Detection OR Localization OR Reconstruction OR (Radiotherapy Planning)). Then the duplicate papers will be removed. We set the qualified publication date to 2011. The remained papers will go through qualitative synthesis and quantitative synthesis. The summary of the selection process is shown in Figure 5.

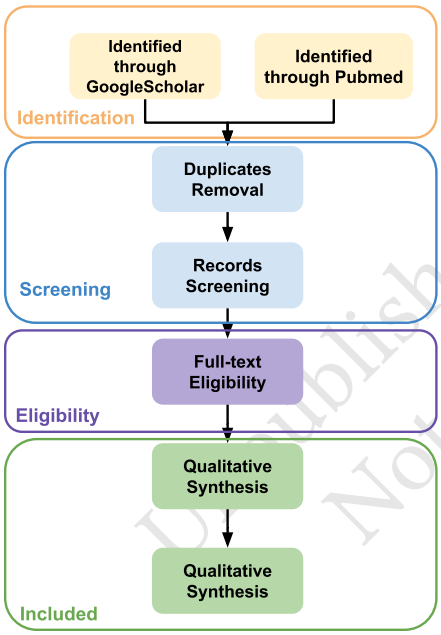

Figure 5: Flows of information through the different phases of a systematic review.

By reviewing content, analyzing common points and comparing difference of these papers, we hope that we can inspire our target readers to (i)have a better understanding of RF, (ii) learn how to formulate their research problems as RF problems. For the next two sections, we will first prepare the readers with basic knowledge of RL. Then we will show how to apply RF in anatomical landmark detection. Those readers who have already been familiar with RF algorithms could directly go to the application section.

2 REINFORCEMENT LEARNING BASICS

In this subsection, we provide a list of terminologies that frequently appear in RF papers. Some terminologies may appear in definitions of other terminologies before they are defined.

- *Action (A):* An action (a) is the way that an agent interacts with the environment. A includes all possible actions that an agent could perform. By taking action, the agent can transfer from the current state to the new state. The agent takes action according to the policy, observing the current state.
- *Agent:* Agents are the models we attempt to build that interact with the environment and take actions.
- *Environment:* The content that the agent is interacting with is called the environment. While providing feedback after the agent takes action, the environment itself is also changing.
- *State (S):* A state (s) is a frame of an environment. S includes all states that an agent will go through.
- *Reward (R):* A positive reward (r) means an increase possibility of achieving the goal, while a negative reward means the decreased possibility. R includes all the possible reward values the environment may feed back to the agent.
- *Episode:* If an agent has gone through all the states from the initial state to the terminal state, we say this agent has finished the episode.
- *Transition probability* $P(s'|s,a)$*:* $P(s'|s,a)$ is the possibility of transiting to state s' from current state s, taking the action a.
- *Policy* $\pi(a|s)$*:* The policy instructs the agent to choose among actions A under the current state.
- *Return (G):* The return is the cumulative discounted future reward. $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \cdots$, where t is the time and γ is the discount factor.
- *State value* $V^\pi(s)$*:* The expected amount of return from current state. $V^\pi(s) = E[G_t|s_t = s]$, where E is the expectation.
- *Action value* $Q^\pi(s,a)$ *(Q value):* The expected amount of return from current state, taking action s. $Q^\pi(s,a) = E[G_t|s_t = s, a_t = a]$
- *Optimal action value* $Q^\star(s,a)$*:* $Q^\star(s,a) = \max_\pi Q^\pi(s_t, a_t)$
- *Agent environment interaction:* Figure 6 shows how the agent is interaction with the environment.

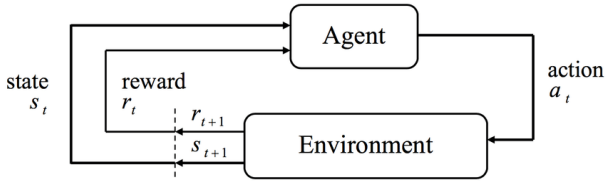

Figure 6: Agent Environment Interaction. Adapted from [Rafati and Noelle 2019].

3 RL IN MEDICAL IMAGE DETECTION[1]

3.1 Landmark Detection

Anatomical landmarks are biological coordinates that can be reallocated repeatedly and precisely on images produced by different imaging modalities - computed tomography (CT), ultrasound(US), magnetic resonance imaging (MRI). The accurate detection of anatomical landmarks is the ground for further medical image analysis tasks. Figure 7 is an example of vocal tract landmarks from the MRI image.

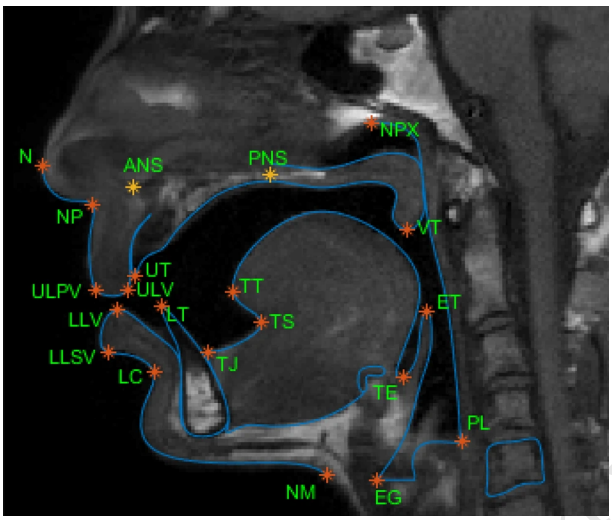

Figure 7: Vocal tract landmarks from MRI image Courtesy of [Eslami et al. 2020].

Many automatic algorithms for anatomical landmark detection have existed long before the attempts of using RF models. However, landmark detection, especially 3D landmarks detection, could be challenging and cause the failure of these algorithms [Ghesu et al. 2019].

- *Scanning-based System:* Sampling complexity of high dimension image could be overwhelming. The false-positive classifier responses could severely influence the result.
- *End-to-End Systems:* Difficult to manage memory during the training process, and the training time is long.
- *Regression-based Systems:* Very hard to train this kind of system and sensitive to the difference in the anatomical range.
- *Atlas-based Systems:* Requires a lot of computational resources.

Moreover, the computation of features and hyper-parameters selection of the system may not be optimal since the involvement of human decisions. The researchers attempted a different paradigm to address this problem - translate the landmark detection tasks as reinforcement learning problems which is the common goal of the papers we reviewed. While most essential and tricky task in these papers, as you can see later, is designing the state space, action space, and reward space before training the models.

[Ghesu et al. 2016] is one of the very first papers that attempted to use RF for anatomical landmarks detection. In an image I, $\vec{p}_{GT}$ denotes the location of anatomical landmark, and $\vec{p}_t$ denotes the location at the current time. State space S is the collection of all possible states $s_t = I(\vec{p}_t)$. Action space A is the collection of all possible actions by which the agent can move to the adjacent position, as illustrated by Figure 8. Reward space R is defined as

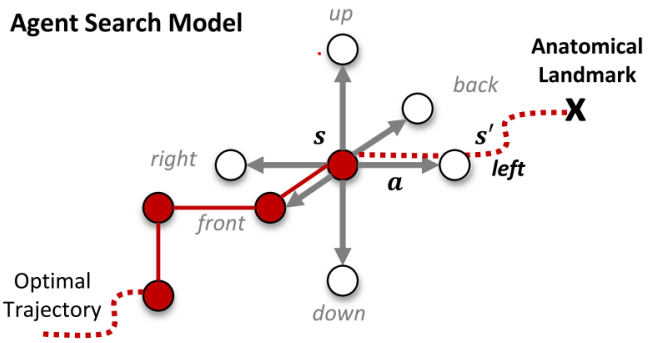

Figure 8: Possible actions of a 3D landmarks detection task. Adapted from [Ghesu et al. 2019].

$||\vec{p}_t - \vec{p}_{GT}||_2^2 - ||\vec{p}_{t+1} - \vec{p}_{GT}||_2^2$ which impels the agent to move closer to the target anatomical landmark. A deep learning model was applied to approximate the state value function. The parameters are updated according to gradient descent, and the error function is equation 1:

$$\hat{\theta}_i = \arg\min_{\theta_i} E_{s,a,r,s'}[(y - Q(s,a;\theta_i))^2] + E_{s,a,r}[V_{s'}(y)], \quad (1)$$

This deep Q learning-based method beat the existing top systems not only in accuracy but also in speed. The design of action, state, and reward spaces in the paper we just discussed became a standard method. So unless particular emphasis, we assume that the rest of the papers we reviewed took the same design.

However, the approach mentioned above is still preliminary. One of the biggest disadvantage that it could not fully use the information at different scale. So a multi-scale deep reinforcement learning method was soon proposed in [Ghesu et al. 2019]. The search for the landmark started from the coarsest scale. Once the search is convergent, the continued work would be started at a finer scale until the search meets the finest scale's convergence criteria. Figure 9 illustrates this non-trivial search process. Where L_d is the scale level in the continuous scale-space L, which can be calculated as equation 2:

$$L_d(t) = \psi_\rho(\sigma(t-1) * L_d(t-1)), \quad (2)$$

where ψ_ρ is the signal operator, and σ is the Gaussian-like smoothing function.

[Alansary et al. 2019] extended the work of Ghesu et al. by evaluating different types of RF agents. He compared the detection results of using DQN, double DQN (DDQN), Double DQN, and duel double DQN (duel DDQN) on three different-modalities dataset -

[1]THIS ARTICLE IS EXCERPTED FROM THE WORK - *REINFORCEMENT LEARNING IN MEDICAL IMAGE ANALYSIS*. IF YOU ARE INTERESTED IN THE FULL ARTICLE, PLEASE CONTACT THE FIRST AUTHOR.

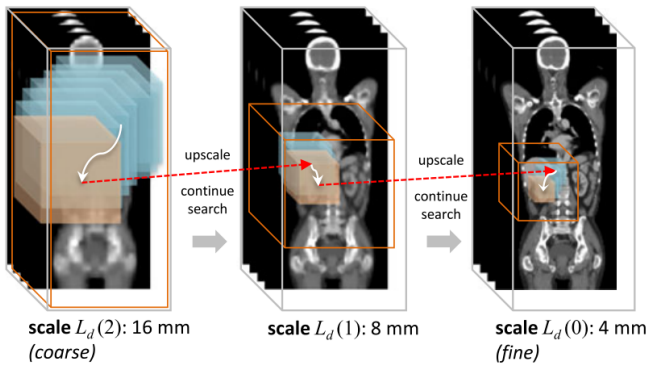

Figure 9: The trajectory of search the anatomical landmark across images of multiple scale-levels. CC BY-NC-SA: Creative Commons Attribution-Noncommercial-ShareAlike. [Ghesu et al. 2019].

fetal US, cardiac MRI, and brain MRI. The result discussion will be covered in the next section.

Rather than detecting a landmark per agent separately, bold attempts have been made by [Vlontzos et al. 2019] to detect multiple landmarks with multiple collaboration agents. With the assumption that the anatomical landmarks have inner correlations with each other, the detection of one landmark could indicate the location of some other landmarks. For the action function approximator in this paper, the collaborative deep Q network (Collab-DQN) was proposed. The weights of the convolutional layers are shared by all the agents, while the fully connected layers for deciding the actions are trained separately per agent. Compared to the methods that trained agents for different landmarks differently, this multi-agent approach reduced 50% detection error using a shorter training time.

The papers we have reviewed above all attempted to learn the action value function $Q^{\pi}(s, a, \theta)$, where θ is the weight of the deep learning approximator. [Abdullah Al and Yun 2020] tried to use an partialpolicy actor-critic (AC) network to solve the problem. The cost function for the value function and policy function update is shown as below:

$$J_V(\omega) = E_{s,a,s'}[(\tau(s,a,s') - V_\omega(s))^2] \quad (3)$$
$$J_V(\omega) = E_{s,a,s'}[-\epsilon(s,a,s')\log_{\pi_\theta}(s,a)],$$

where τ is the temporal difference target (TD target), and ϵ is the temporal difference error (TD error). Experiments were conducted on 3D CT and MR images. This paper asserted that the number of trials and data required was significantly reduced by adopting the partial policy-based actor-critic method compared to the conventional DQN methods. Most importantly this proposed partial policy-based reinforcement learning method (PPRL) converged much faster than the conventional DQN methods and AC methods.

Some other contributions to the RF for anatomical landmarks detection include: estimating the uncertainty of reinforcement learning agent [Browning et al. 2021], reducing the needed time to reach the landmark by using a continuous action space [Kasseroller et al. 2021], localization of modality invariant landmark [Winkel et al. 2020]. Since these papers are less useful in helping readers understand and build their own RF systems for landmark detection, we won't review them in detail here. A conclusion of the article we reviewed in this paper is given in Table 1

3.2 Lesion Detection

Object detection, also called object extraction, is the process of finding out the class labels and locations of target objects in images or videos. It is one of the primary tasks in medical image analysis [Li et al. 2019]. An exemplary detection result can be used as the basis to improve the performance of further tasks like segmentation.

The mainstream approaches for lesion detection nowadays still rely on exhaustive search methods that cost a lot of time and deep learning methods that require a large amount of labeled data. Facing the current challenges and inspired by similar problems in landmarks detection [Ghesu et al. 2016], [Maicas et al. 2017] implemented a deep Q-network (DQN) agent for active breast lesion detection. The states are defined as current bounding box volumes of the 3D DCE-MR images. The reinforcement learning agent could gradually learn the policy to choose among actions to transit, scale the bounding box, and finally localize the breast lesion. Specifically, the action set consists of 9 actions that can translate the bounding box forward or backward along the x, y, z-axis, scale up or scale down the bounding box, and trigger the terminal state.

To further evaluate the effectiveness of applying reinforcement learning on lesion detection with limited data, using DQN as the agent to localize brain tumors with very small training data was attempted by [Stember and Shalu 2020], [Stember and Shalu 2021]. Different from [Maicas et al. 2017], the brain MR data are 2D slices. The environment is defined as the 2D slices overlaid with gaze plots viewed by the radiologist. Instead of using the bounding box, the states are the gaze plots the agent located. Three actions - moving anterograde, not moving, moving retrograde would help the agent transfer to the next state. If the agent moves toward the lesion, it will receive a positive reward, otherwise a negative one. If the agent stays still, it will receive a relatively large positive reward within the lesion area or a rather large penalization otherwise. The experiment results showed that reinforcement learning models could work as robust lesion detectors with limited training data, reduce time consumption, and provide some interpretability.

Also addressing the lack of labeled training data, [Pesce et al. 2019] exploited visual attention mechanisms to learn from a combination of weakly labeled images (only class label) and a limited number of fully annotated X-ray images. This paper proposed convolutional networks with attention feedback (CONAF) architecture and a recurrent attention model with annotation feedback (RAMAF) architecture. The RAMAF model can only observe one part of the image, which is defined as a state at a glimpse. The reinforcement learning agent needs to learn the policy to take a sequence of glimpses and finally locate the lesion site within the shortest time. Each glimpse consists of two image patches sharing the same central point, and the length of the glimpse sequence is fixed to be 7. The rewards will be decided according to (i) if the image is classified correctly; (ii) if the central point of a glimpse is within

Table 1: Summary of RF in landmark detection

Article	Year	Modality	RF Taxonomy	Highlights
[Abdullah Al and Yun 2020]	2020	CT & MRI	Policy gradient, Model free	Faster convergence, less data and trails required
[Alansary et al. 2019]	2019	Fetal head US & Cardiac MRI & Brain MRI	Value based, Model free	Deep Q Network (DDQN, duel DQN, duel DDQN)
[Vlontzos et al. 2019]	2019	Brain MRI & Cardiac MRI & Brain US	Value based, Model free	Multi-agent
[Ghesu et al. 2016]	2016	MRI, Head-neck CT, Cardiac US	Value based, Model free	First formulation of action space, reward space, state space
[Ghesu et al. 2019]	2019	Large 3D CT	Value based, Model free	Multiscale searching

the labeled bounding box. RAMAF achieved a localization performance of detecting 82% of overall bounding boxes with a much faster detection speed than other state-of-the-art methods.

More than detecting lesions in static medical images (2D or 3D), the reinforcement-learning-based system can also track the lesions frame by frame continuously. [Luo et al. 2019] proposed a robust RL-based framework to detect and track plaque in Intravascular Optical Coherence Tomography (IVOCT) images. Despite the pollution problem of speckle-noise, blurred plaque edges, and diverse intravascular morphology, the proposed method achieved accurate tracking and has strong expansibility.

Three different modules are included in the proposed framework. The features are extracted and encoded first by the enoding feature module. Then the information of scale and location of the lesion is provided by the localization and identification module. Another function of this module is preventing over-tracking. The most important module is the spatial-temporal correlation RL module. Nine different actions are different, including eight transformation actions and one stop action. The state S is defined as three-tuples: $S = (E, HL, HA)$. Here, the E represents the encoded output features from the FC1 layer. HL is the collection of recent locations and scales. HA represents the recent ten sets of actions. 8000 IVOCT images were used to evaluate the framework. With a strict standard ($IOU > 0.9$), the RL module could improve the performance of plaque tracking both frame-level and plaque-level. Table 2 summarized RF papers in lesion detection.

3.3 Organ/ Anatomical Structure Detection

Besides detecting lesions, reinforcement learning can also be applied in organ detection. [Navarro et al. 2020] designed a deep Q-learning agent to locate various organs in 3D CT scans. The state is defined as voxel values within the current 3D bounding box. Eleven actions, including six translation actions, two zooming actions, and three scaling actions, make sure that the bounding box can move to any part of the 3D scan. The agent is rewarded if an action improves the intersection over union score (IOU). Seventy scans were used for training, and 20 scans were used for testing on seven different organs: pancreas, spleen, liver, lung (left and right), and kidney (left and right). This proposed method achieved a much faster speed than the region proposal and the exhaustive search methods and led to an overall IOU score of 0.63.

[Zhang et al. 2021] managed to detect and segment the vertebral body (VB) simultaneously. The sequence correlation of the VB is learned by a soft actor-critic (SAC) RF agent to reduce the background interference. The proposed framework consists of three modules: Sequential Conditional Reinforcement Learning network (SCRL), FC-ResNet, and Y-net. The SCRL learns the correlation and gives the attention region. The FC-ResNet extracts the low-level and high-level features to determine a more precise bounding box according to the attention region. At the same time, the segmentation result is provided by the Y-net. The state of the RF agent is determined by a combination of the image patch, feature map, and region mask. And the reward is designed according to the change of attention-focusing accuracy to elicit the agent to achieve a better detection performance. This proposed approach accomplished an average of 92.3% IOU on VB detection and an average of 91.4% Dice on VB segmentation. Figure 10 shows the process of detecting multiple vertebral bodies (VBs) simultaneously by SCRL.

The research of [Zheng et al. 2021] was the first attempt to use the multi-agent RF in prostate detection. Two DQN agents locate the lower-left and upper-right corners of the bounding box while sharing knowledge according to the communication protocol [Foerster et al. 2016]. The final location of the prostate is searched with a coarse-to-fine strategy to speed up the search process and improve the detection accuracy. In more detail, the agents first search on the coarsest scale to draw a big bounding box and gradually move to a finer scale to generate a smaller and more accurate bounding box to detect the prostate. Compared to the single-agent strategy (63.15%), this multi-agent framework achieved a better average score of 80.07% in IOU. Table 3 summarized RF papers in organ detection.

4 CONCLUSIONS AND FUTURE PROSPECTIVES

From our review, we have witnessed the success of some researchers in using the RF methods for medical image analysis, and we are

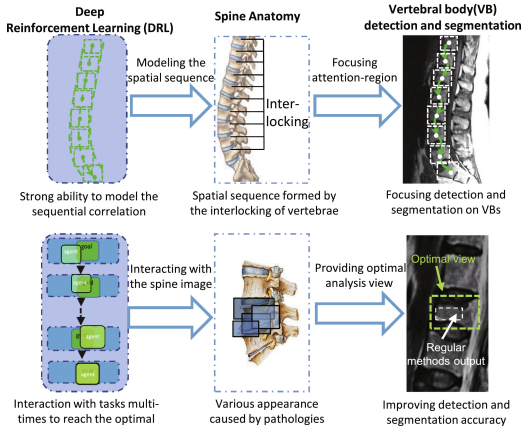

Figure 10: Detecting multiple VBS simultaneously by exploiting the spatial sequence correlation knowledge. Courtesy of [Zhang et al. 2021].

optimistic about the further improvements of the current methods. The application of RF creates a new paradigm of solving the present obstacles as behavioral learning problems. The RF method has shown its unique advantage in finding the anatomical landmarks accurately while providing the optimized search path, which is different from conventional supervised and unsupervised learning methods. Moreover, RF reduced the time cost of finding the solution, especially for large 3D images, and the chance of missing the landmark point in detection.

However, limitations still exist in current research. It is very tricky to design the action space, state-space, and reward space. It highly relies on the researcher's experience, and bad design may cause the model never to converge. There is no standard for choosing between the RF models. It seems that many researchers randomly pick an RF model without explaining the reason for choosing it. It might be tough to reproduce the results. Reinforcement learning itself is a highly complex field. The few pages of description in a paper are not enough for the readers to re-implement the method. Moreover, because of privacy concerns, the data and codes of these research are rarely available to the public. Designing RF systems for medical analysis tasks requires the researcher to have strong background knowledge in computer science, statistics, and the medical domain, preventing many medical researchers from applying RF models in their research. We hope researchers can address these points in future studies to promote faster RF growth for medical image analysis.

DISCLOSURE

The authors are not aware of any affiliations, memberships, funding, or financial holds that might be perceived as affecting the objectivity of this review.

Table 2: Summary of RF in lesion detection

Reference	Site	Modality	RL Agent	States	Actions	Highlight
Luo, G., et al. (2019).	Heart	IVOCT	Action-decsion Net (ADNet)	Spatial-temporal locations correlation information	9	Tracked plaque continuously and accurately frame-by-frame on IVOCT images
Maicas, G., et al. (2017).	Breast	3D DCE-MRI	DQN	Current bounding volume	9	1.78 times faster than the exhaustive methods. Small labeled training set needed
Pesce, E., et al. (2019).	Lung	X-ray	REINFORCE	The part of the image observed by the glimpse	Continuous	Trained the model only with weakly-labeled image together with few annotated image
Stember, J. and H. Shalu (2020)	Brain	2D MRI	DQN	The gaze plot the agent locate	3	Detect the lesion accurately with limited training data
Zheng, C., et al. (2021).	Prostate	MRI	DQN	Voxel values contained in the bounding box	4	The first research in using multi-agent reinforcement learning for prostate detection

Table 3: Summary of RF in organ detection

Reference	Site	Modality	RL Agent	States	Actions	Highlight
Navarro, F., et al. (2020).	Lung, Kidney, Liver, Spleen, Pancreas	CT	DQN	Voxel value of the current bounding box	11	Designed actions that work best for organ detection Achieved the detection and segmentation at the same time. Solved the
Zhang, D., et al. (2021).	Vertebral Body	MRI	Soft Actor-Critic (SAC)	Combination of the image patch, feature map and region mask.	4	Unbalanced-experiences, Reward-confusion and Fixed-exploration problem to build a more robust RL model.

Appendices

A PSEUDO-CODE OF RF ALGORITHMS

A.1 Q-learning

Result: Off-policy control for estimating $\pi \simeq \pi_*$
Parameters: step size $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q(s, a)\ \forall s \in \mathcal{S}^+, a \in \mathcal{A}(s)$ arbitrarily.
Set $Q(terminal, \cdot) = 0$
for *each episode* **do**
  Initialize S;
  for $step = 0$ **to** T **do**
    Choose A from S using policy derived from Q (e.g., ϵ-greedy);
    Take action A, observe R, S';
    $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$;
    $S \leftarrow S'$;
  end
end

A.2 Deep Q-learning DQN

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

A.3 Actor Critic (AC)

Randomly initialize critic network $V_\pi^U(s)$ and actor network $\pi^\theta(s)$ with weights U and θ
Initialize environment E
for** episode = 1, M **do
  Receive initial observation state s_0 from E
  for** t=0, T **do
    Sample action $a_t \sim \pi(a|\mu, \sigma) = \mathcal{N}(a|\mu, \sigma)$ according to current policy
    Execute action a_t and observe reward r and next state s_{t+1} from E
    Set TD target $y_t = r + \gamma \cdot V_\pi^U(s_{t+1})$
    Update critic by minimizing loss: $\delta_t = \left(y_t - V_\pi^U(s_t)\right)^2$
    Update actor policy by minimizing loss:
      $Loss = -log\left(\mathcal{N}(a \mid \mu(s_t), \sigma(s_t))\right) \cdot \delta_t$
    Update $s_t \leftarrow s_{t+1}$
  end for
end for

REFERENCES

Abdullah Al, W. and Yun, I. D. [2020]. Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images, *IEEE Transactions on Medical Imaging* **39**(4): 1245–1255.

Alansary, A., Oktay, O., Li, Y., Folgoc, L. L., Hou, B., Vaillant, G., Kamnitsas, K., Vlontzos, A., Glocker, B., Kainz, B. and Rueckert, D. [2019]. Evaluating reinforcement learning agents for anatomical landmark detection, *Medical Image Analysis* **53**: 156–164.

Browning, J., Kornreich, M., Chow, A., Pawar, J., Zhang, L., Herzog, R. and Odry, B. L. [2021]. Uncertainty aware deep reinforcement learning for anatomical landmark detection in medical images, *in* M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng and C. Essert (eds), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, Cham, pp. 636–644.

Eslami, M., Neuschaefer-Rube, C. and Serrurier, A. [2020]. Automatic vocal tract landmark localization from midsagittal mri data, *Scientific Reports* **10**(1): 1–13.

Foerster, J. N., Assael, Y. M., de Freitas, N. and Whiteson, S. [2016]. Learning to communicate with deep multi-agent reinforcement learning.

Ghesu, F. C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J. and Comaniciu, D. [2016]. An artificial agent for anatomical landmark detection in medical images, *in* S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal and W. Wells (eds), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, Springer International Publishing, Cham, pp. 229–237.

Ghesu, F.-C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J. and Comaniciu, D. [2019]. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(1): 176–189.

Kasseroller, K., Thaler, F., Payer, C. and Štern, D. [2021]. Collaborative multi-agent reinforcement learning for landmark localization using continuous action space, *in* A. Feragen, S. Sommer, J. Schnabel and M. Nielsen (eds), *Information Processing in Medical Imaging*, Springer International Publishing, Cham, pp. 767–778.

Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P. and Zeng, Z. [2019]. Clu-cnns: Object detection for medical images, *Neurocomputing* **350**: 53–59.

Luo, G., Dong, S., Wang, K., Zhang, D., Gao, Y., Chen, X., Zhang, H. and Li, S. [2019]. A deep reinforcement learning framework for frame-by-frame plaque tracking on intravascular optical coherence tomography image, *in* D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan (eds), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, Cham, pp. 12–20.

Maicas, G., Carneiro, G., Bradley, A. P., Nascimento, J. C. and Reid, I. [2017]. Deep reinforcement learning for active breast lesion detection from dce-mri, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, Springer, pp. 665–673.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. and Group, P. [2009]. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement, *PLoS medicine* **6**(7): e1000097.

Navarro, F., Sekuboyina, A., Waldmannstetter, D., Peeken, J. C., Combs, S. E. and Menze, B. H. [2020]. Deep reinforcement learning for organ localization in ct, *Medical Imaging with Deep Learning*, PMLR, pp. 544–554.

Pesce, E., Joseph Withey, S., Ypsilantis, P.-P., Bakewell, R., Goh, V. and Montana, G. [2019]. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks, *Medical Image Analysis* **53**: 26–38.

Rafati, J. and Noelle, D. C. [2019]. Learning representations in model-free hierarchical reinforcement learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 10009–10010.

Sallab, A. E., Abdou, M., Perot, E. and Yogamani, S. [2017]. Deep reinforcement learning framework for autonomous driving, *Electronic Imaging* **2017**(19): 70–76.

Sharma, A. R. and Kaushik, P. [2017]. Literature survey of statistical, deep and reinforcement learning in natural language processing, *2017 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, pp. 350–354.

Shen, D., Wu, G. and Suk, H.-I. [2017]. Deep learning in medical image analysis, *Annual Review of Biomedical Engineering* **19**: 221–248.

Stember, J. N. and Shalu, H. [2021]. Reinforcement learning using deep q networks and q learning accurately localizes brain tumors on mri with very small training sets.

Stember, J. and Shalu, H. [2020]. Deep reinforcement learning to detect brain lesions on mri: a proof-of-concept application of reinforcement learning to medical images.

Sutton, R. S. and Barto, A. G. [2018]. *Reinforcement learning: An introduction*, MIT press.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P. et al. [2019]. Grandmaster level in starcraft ii using multi-agent reinforcement learning, *Nature* **575**(7782): 350–354.

Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D. and Kainz, B. [2019]. Multiple landmark detection using multi-agent reinforcement learning, *in* D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan (eds), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, Cham, pp. 262–270.

Wang, T., Lei, Y., Fu, Y., Wynne, J. F., Curran, W. J., Liu, T. and Yang, X. [2021]. A review on medical imaging synthesis using deep learning and its clinical applications, *Journal of Applied Clinical Medical Physics* **22**(1): 11–36.

Wilson, S. M. and Anagnostopoulos, D. [2021]. Methodological guidance paper: The craft of conducting a qualitative review, *Review of Educational Research* p. 00346543211012755.

Winkel, D. J., Weikert, T. J., Breit, H.-C., Chabin, G., Gibson, E., Heye, T. J., Comaniciu, D. and Boll, D. T. [2020]. Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation, *European Journal of Radiology* **126**: 108918.

Zhang, D., Chen, B. and Li, S. [2021]. Sequential conditional reinforcement learning for simultaneous vertebral body detection and segmentation with modeling the spine anatomy, *Medical Image Analysis* **67**: 101861.

Zheng, C., Si, X., Sun, L., Chen, Z., Yu, L. and Tian, Z. [2021]. Multi-agent reinforcement learning for prostate localization based on multi-scale image representation, *in* H. Lu, S. Mu and S. Nakashima (eds), *International Symposium on Artificial Intelligence and Robotics 2021*, Vol. 11884, International Society for Optics and Photonics, SPIE, pp. 487 – 494.