


Please cite the Published Version

Ahmad, Farooq, Zhang, Xinfeng, Tang, Zifang, Sabah, Fahad, Azam, Muhammad and Sarwar, Raheem  (2025) Deep deterministic policy gradients with a self-adaptive reward mechanism for image retrieval. The Journal of Supercomputing, 81 (1). 336 ISSN 0920-8542

DOI: <https://doi.org/10.1007/s11227-024-06764-9>

Publisher: Springer

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/637723/>

Usage rights:  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Additional Information: This is an open access article which first appeared in The Journal of Supercomputing

Data Access Statement: Data will be made available on request. Code will be made available/public on request.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



Deep deterministic policy gradients with a self-adaptive reward mechanism for image retrieval

Farooq Ahmad¹ · Xinfeng Zhang¹ · Zifang Tang² · Fahad Sabah² ·
Muhammad Azam³ · Raheem Sarwar⁴

Accepted: 22 November 2024
© The Author(s) 2024

Abstract

Traditional image retrieval methods often face challenges in adapting to varying user preferences and dynamic datasets. To address these limitations, this research introduces a novel image retrieval framework utilizing deep deterministic policy gradients (DDPG) augmented with a self-adaptive reward mechanism (SARM). The DDPG-SARM framework dynamically adjusts rewards based on user feedback and retrieval context, enhancing the learning efficiency and retrieval accuracy of the agent. Key innovations include dynamic reward adjustment based on user feedback, context-aware reward structuring that considers the specific characteristics of each retrieval task, and an adaptive learning rate strategy to ensure robust and efficient model convergence. Extensive experimentation with the three distinct datasets demonstrates that the proposed framework significantly outperforms traditional methods, achieving the highest retrieval accuracy having 3.38%, 5.26%, and 0.21% improvement overall as compared to the mainstream models over DermaMNIST, PneumoniaMNIST, and OrganMNIST datasets, respectively. The findings contribute to the advancement of reinforcement learning applications in image retrieval, providing a user-centric solution adaptable to various dynamic environments. The proposed method also offers a promising direction for future developments in intelligent image retrieval systems.

Keywords Image retrieval · Reinforcement learning · Self-adaptive reward mechanism · Deep deterministic policy gradients · Adaptive learning rate

Xinfeng Zhang, Zifang Tang, Fahad Sabah, Muhammad Azam and Raheem Sarwar have contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

Image retrieval systems have undergone significant evolution, transitioning from traditional keyword-based search methods to more advanced approaches that leverage deep learning for feature extraction and similarity matching. Early methods like the bag-of-words model and scale-invariant feature transform (SIFT) descriptors focused on extracting handcrafted features to match image content. These techniques struggled to capture the high-level semantics of visual data, leading to limited retrieval accuracy and user satisfaction.

In recent years, convolutional neural network (CNN) has revolutionized the field by automatically learning hierarchical feature representations from large datasets. CNN enables more precise image retrieval by generating deep feature vectors that better represent the content and semantics of images, improving the system's ability to retrieve relevant results based on visual similarity. However, static deep learning models have limitations when user preferences change dynamically or when the relevance of the content evolves over time. To address these issues, researchers have begun exploring reinforcement learning (RL) as a way to introduce adaptability and interaction-based learning into image retrieval systems. A general framework for RL is shown in Fig. 1

RL, particularly in combination with deep neural networks, has proven effective in various domains such as gaming, robotics, and natural language processing. RL algorithms enable systems to adapt their strategies based on real-time feedback, making them ideal for tasks where user preferences and contextual relevance are constantly shifting. The application of RL to image retrieval is still in its nascent stages but shows great promise. Existing research, such as the work of Khamaj et al. [1] and Liang et al. [2], has demonstrated RL's potential for improving user interactions and adaptability in various fields.

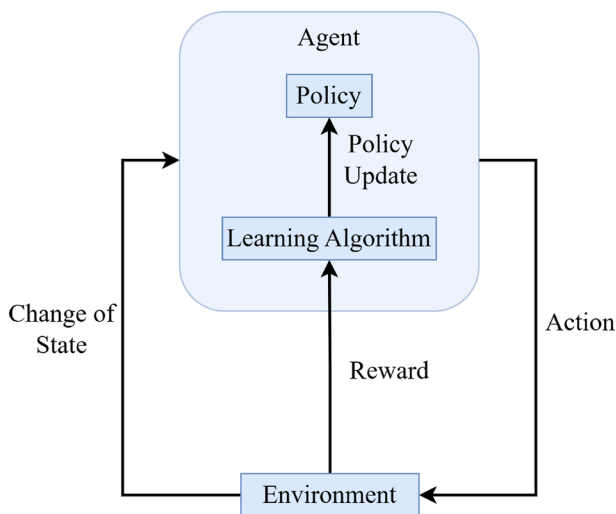


Fig. 1 Reinforcement learning framework

Despite this progress, many RL-based image retrieval systems lack mechanisms that can effectively interpret and respond to user feedback and contextual changes. The integration of advanced reward mechanisms within RL frameworks is an emerging research area aimed at addressing these limitations. By incorporating dynamic, user-driven reward systems, it is possible to improve the relevance and personalization of retrieval results, ensuring that the system remains responsive to user needs over time.

In the context of image retrieval, the self-adaptive reward mechanism contributes to more personalized and effective retrieval outcomes. It helps in prioritizing images that are more likely to meet user expectations, thereby enhancing user satisfaction and engagement with the retrieval system [3]. Moreover, it supports continuous improvement in retrieval performance as the system operates in real-world scenarios. Operating in real-world scenarios entails encountering diverse user preferences and evolving content dynamics. The adaptive nature of the mechanism equips the retrieval system to respond effectively to these challenges. It adapts dynamically to changes in user behavior and content relevance, ensuring robust performance across varying conditions and contexts.

Medical image analysis plays a crucial role in assisting physicians with both qualitative and quantitative assessments of lesions and anatomical structures, thereby enhancing the accuracy and reliability of medical diagnoses and prognoses. Traditionally, these tasks were labor-intensive and susceptible to inefficiencies and biases when performed solely by experienced physicians or medical physicists. During the past decade, there has been a notable surge in the application of machine learning methods to streamline and automate medical image analysis. Despite the widespread adoption of supervised and unsupervised learning models, the utilization of RL in this domain remains relatively limited [4].

Image retrieval systems have evolved from simple keyword-based searches to complex models leveraging deep learning for feature extraction and similarity matching. Traditional methods, such as bag-of-words and SIFT descriptors, have laid the foundation for image retrieval. Recent advancements have employed convolutional neural network (CNN) to extract deep features, significantly improving retrieval performance. RL has introduced a new dimension to image retrieval, enabling dynamic and adaptive search strategies. However, the potential benefits of integrating RL with advanced reward mechanisms in image retrieval are not fully realized. Research in this area can lead to systems that not only deliver more relevant results but also continuously improve their performance based on user interactions and contextual changes [5].

In the realm of image retrieval, several challenges hinder the effectiveness of traditional systems. Firstly, dynamic user preferences pose a significant obstacle, as users often have changing interests and needs influenced by personal experiences, trends, or seasonal factors. This variability can lead to irrelevant results when systems fail to adapt. Additionally, contextual changes, such as current events or specific situational requirements, further complicate the retrieval process, as the

relevance of images can shift dramatically. Many existing systems also struggle with effectively interpreting and integrating user feedback, limiting their ability to learn from interactions and improve over time. Moreover, the evolution of content relevance necessitates continuous reassessment of retrieved results, as certain images may gain or lose significance over time. Finally, achieving high levels of personalization in retrieval outcomes remains challenging, particularly when systems must cater to diverse user preferences and context-specific needs simultaneously.

In this research work, we propose a novel image retrieval framework that combines deep deterministic policy gradient (DDPG) with a self-adaptive reward mechanism (SARM). This approach aims to address the challenges posed by evolving user preferences and varying contextual factors. The framework interprets both user feedback and contextual cues to dynamically adjust the agent's retrieval strategies, fostering continuous learning and improvement in retrieval performance.

1.1 Objectives

1. **Dynamic reward adjustment based on user feedback** Develop a reward adjustment algorithm that dynamically modifies the reward values based on user feedback to ensure that the RL agent learns to prioritize user satisfaction and relevance in image retrieval.
2. **Context-aware reward structuring** Introduce a context-aware component to the reward mechanism that considers the specific characteristics of the retrieval task, such as image complexity, diversity, and user-specific preferences, to provide a more nuanced reward structure.
3. **Adaptive learning rate for efficient convergence** Implement an adaptive learning rate strategy within the DDPG framework that adjusts the learning rate based on the agent's performance and reward trends, facilitating faster and more stable convergence of the RL model.

1.2 Motivations

Managing image complexity in retrieval tasks

Image retrieval, particularly in specialized fields such as healthcare, often involves handling complex images with varying levels of detail, such as medical images with intricate features. Current retrieval systems struggle to accurately capture and interpret this complexity, leading to suboptimal results. A RL-based framework that incorporates contextual factors, such as image complexity and user-specific needs, is essential to improve the accuracy and relevance of retrieval outcomes.

Addressing data imbalance in image datasets

Image datasets, especially those used in niche domains like medical or e-commerce applications, often suffer from imbalanced class distributions. This imbalance

can lead to biased retrieval results, where frequently occurring image types are prioritized at the expense of rarer, but equally important, images. Developing a dynamic learning approach that adjusts based on varying data distributions is critical to ensure balanced and unbiased retrieval.

Adapting to dynamic user preferences and feedback

User preferences can vary widely, and these preferences can evolve over time as they interact with image retrieval systems. Traditional static retrieval models fail to capture this dynamic aspect of user behavior. Therefore, a self-adaptive reward mechanism that continuously learns from user feedback is necessary to personalize retrieval experiences and improve satisfaction by dynamically prioritizing relevant results.

Optimizing learning for efficient and stable convergence

RL models can face challenges in achieving stable and efficient learning, particularly when dealing with complex, high-dimensional tasks such as image retrieval. An adaptive learning rate mechanism that adjusts based on the agent's performance is crucial to ensuring faster convergence and more robust model training, leading to a more responsive and efficient retrieval system.

1.3 Contributions

This research makes several significant contributions to the field of image retrieval using RL:

Development of an RL framework The primary contribution lies in the design and implementation of an RL framework tailored specifically for image retrieval tasks. By integrating deep learning models, particularly convolutional neural network (CNN), with RL algorithms such as deep deterministic policy gradient (DDPG), the framework enables the autonomous learning of optimal image retrieval strategies based on user interactions and feedback.

Enhanced adaptability and accuracy Unlike traditional methods that rely on static feature extraction and heuristic-based approaches, the proposed framework enhances adaptability by dynamically learning and adjusting its retrieval strategy. This adaptability leads to improved accuracy in selecting relevant images from large-scale datasets, thereby enhancing the overall precision and effectiveness of the retrieval system.

Integration of user feedback mechanisms A key contribution is the integration of a self-adaptive reward mechanism within the RL framework. This mechanism enables the system to effectively incorporate user feedback into the learning process, facilitating personalized and context-aware image retrieval. By dynamically adjusting rewards based on user interactions, the system improves its responsiveness and relevance to user preferences over time.

Empirical evaluation and validation The research includes comprehensive empirical evaluations conducted on standard benchmark datasets, such as MNIST, and potentially on larger and more complex datasets. These evaluations assess the

performance, efficiency, and scalability of the proposed RL-based image retrieval system compared to traditional methods. The results provide empirical evidence of the system's effectiveness and demonstrate its potential for real-world applications.

Practical applications and future directions Beyond theoretical advancements, this research explores practical applications of the proposed framework across various domains, including healthcare, e-commerce, and digital libraries. By demonstrating its applicability in diverse settings, the research contributes to expanding the scope and impact of RL techniques in image retrieval.

Contribution to research and development Lastly, this research contributes to the broader research community by advancing the state-of-the-art in image retrieval methodologies. It provides insights into the integration of deep learning with RL for complex tasks, paving the way for future research directions in autonomous, adaptive systems for image analysis and retrieval.

In summary, the contributions of this research extend beyond technical advancements to encompass practical applications, empirical validations, and implications for the future development of intelligent image retrieval systems leveraging RL.

1.4 Structure of this paper

This paper is structured to provide a comprehensive understanding of the proposed framework for image retrieval based on RL. The structure unfolds as follows.

Section 1 introduces the paper, outlining the objectives, motivations, contributions, and overall structure of this paper. Section 2 presents the related work, summarizing recent advancements in reinforcement learning (RL), image retrieval, and user feedback mechanisms. This section is divided into subsections, covering topics such as rewards adjustment, incorporating user feedback, contextual adaptation, and enhanced learning and adaptability. Section 3 provides the problem formulation. The elements of this section establish a comprehensive mathematical foundation for the proposed approach. In Sect. 4, the proposed methodology is discussed, focusing on the self-adaptive reward mechanism with its dynamic adjustment factor and context factor. Section 5, titled experimental setup, includes descriptions of the datasets used, along with information on hyperparameters, data transformation and loading, neural network model and features extraction, actor and critic networks, training and validation loop, and evaluation metrics. Section 6 provides the results and discussion, offering a comparison of DDPG-SARM with state-of-the-art (SOTA) methods and including an ablation study to assess model robustness. Section 7 concludes the paper with a summary of key findings, and Sect. 8 outlines possible future works.

2 Related work

In this section, we review the existing literature related to key aspects of our proposed method, including rewards adjustment, incorporating user feedback, contextual adaptation, enhanced learning and adaptability, and benefits in image retrieval.

Each of these components contributes to the development of more sophisticated and personalized image retrieval systems.

2.1 Rewards adjustment

Dynamic reward adjustment enhances the DDPG agent's capacity to learn by allowing rewards to evolve based on recent interactions and contextual information. This ongoing adjustment enables the agent to quickly adapt its retrieval strategies, improving its responsiveness to changing user preferences or dataset characteristics. As a result, the agent's learning process becomes more aligned with user needs and the complexity of the task at hand.

Lillicrap et al. [6] successfully demonstrated a model-free approach using DDPG, solving more than 20 physics-based tasks, including cartpole swing and dexterous manipulation. However, DDPG, like many reinforcement learning (RL) methods, requires a significant number of training episodes. Despite this, it remains an integral component for addressing various RL challenges due to its adaptability. Zhao et al. [7] introduced continuous-time RL methods, showcasing the performance of policy gradient methods such as trust region policy optimization/proximal policy optimization (TRPO/PPO) in the continuous setting. Their research emphasizes the importance of policy adaptation to continuous environments, which resonates with our aim to enhance dynamic reward adjustment in image retrieval.

Viswanadhapalli et al. [8] applied DDPG in control systems, using reward shaping for reference tracking in flexible manipulators. This work underscores the value of incorporating domain-specific constraints into reward mechanisms, similar to our approach in tailoring rewards for the dynamic nature of image retrieval tasks. Furthermore, recent advances, such as the reward-adaptive RL method by [9], introduce hybrid policy gradients that optimize multiple criteria simultaneously. The ability to dynamically prioritize different reward components is crucial in environments like image retrieval, where multiple user-specific and contextual factors come into play.

In dynamic reward frameworks, as seen in the multi-reward architecture (MRA) proposed by Xu et al. [10], learning from multiple sub-reward branches leads to more granular decision-making. In our image retrieval framework, dynamically adjusting reward structures based on ongoing feedback will provide similar benefits, allowing the system to better align with evolving user preferences. Ultimately, dynamic rewards adjustment equips the DDPG agent with the flexibility to prioritize actions that maximize positive outcomes based on current user feedback. This adaptability is essential for ensuring that the image retrieval process remains relevant, personalized, and capable of delivering results that align with user expectations. The proposed dynamic reward adjustment mechanism is particularly beneficial in the context of image retrieval, where static approaches often fall short of capturing nuanced user preferences and evolving task conditions. By continually updating the reward structure based on real-time interactions and contextual factors, the retrieval system is able to offer more tailored and satisfying results.

The exploration of dynamic reward adjustment in RL has shown promise across various domains, such as robotics and gaming [11]. However, its application to image retrieval remains relatively underexplored. This research addresses that gap by proposing a novel framework that leverages dynamic reward adjustments to enhance the relevance and accuracy of image retrieval tasks, aligning the system's learning process with user-driven objectives and task-specific nuances.

2.2 Incorporating user feedback

Interpreting user interactions with the retrieval system, such as clicks on retrieved images or other forms of feedback, to determine the quality of retrieval outcomes. Positive feedback, indicating satisfaction with the retrieved images, leads to higher rewards, while negative feedback adjusts the rewards downward. Incorporating user feedback effectively is central to the self-adaptive reward mechanism designed for the image retrieval system. This mechanism relies on interpreting user interactions with the retrieval system, which can include actions such as clicks on retrieved images or explicit feedback provided by the user. These interactions serve as valuable signals that help gauge the quality and relevance of the retrieved images from the user's perspective.

When a user interacts positively with the retrieval system, such as by clicking on retrieved images or indicating satisfaction through explicit feedback, the mechanism responds by assigning higher rewards. This positive reinforcement encourages the agent to prioritize and reinforce actions that lead to satisfying retrieval outcomes, aligning more closely with user preferences and expectations. Conversely, negative interactions or feedback from the user result in adjustments to the rewards provided. This could occur when a user ignores or expresses dissatisfaction with the retrieved images. Lower rewards in response to negative feedback prompt the agent to reassess its strategies and prioritize alternative actions that may better meet user expectations or retrieval needs. In this research work instead of real user feedback, the correctness of the RL agent's action (classification) is used to simulate feedback.

Liang et al. [2] introduced an innovative application of deep reinforcement learning (DRL) for droplet routing on digital microfluidic biochips (DMFBs), which automate biochemical protocols. DMFBs face challenges with electrode degradation over time, which can disrupt droplet transport and compromise bioassay accuracy. By framing droplet routing as a reinforcement learning problem, the authors trained neural network policies to adapt to electrode conditions, ensuring reliable fluidic operations. Their RL-based solution, validated on both real and simulated DMFBs, proves effective across various chip sizes and is computationally feasible on devices like the Raspberry Pi 4, showing promise for time-sensitive bioassays.

Khamaj et al. proposed a modern method to enhance user experiences through RL and a deep Q network (DQN). Using these techniques, the objectives are to optimize user interactions and increase engagement, satisfaction, and task completion rates. Traditional user interfaces offer a common experience due to their impersonal and inflexible nature, which limits the potential for higher engagement and satisfaction in the absence of real-time changes based on individual preferences and behaviors. To address this issue, the study introduces an intelligent system that continuously learns and adapts to user interactions. This innovative approach combines RL and DQN to incrementally adjust user interfaces. Unlike conventional methods, the proposed model adapts by using well-established, high-reward moves along with the development of new strategies through an exploration–exploitation mechanism. Timestamped data fields such as Event-Type, contentId, personId, sensorId, and timestamp provide a comprehensive understanding of user behavior, enabling detailed and nuanced adjustments to the interface [1].

2.3 Contextual adaptation

The self-adaptive reward mechanism considers both user feedback and broader contextual factors in the image retrieval task. This includes evaluating the relevance of retrieved images to specific contexts, such as thematic alignment and situational appropriateness, which helps the agent refine retrieval strategies. By dynamically adapting rewards based on these contextual cues, the mechanism ensures responsiveness to evolving user preferences and trends, optimizing retrieval performance over time. This approach enables the agent to maintain relevance and adapt to changing conditions, enhancing the utility of its retrieval outcomes.

Intelligent manufacturing and agent-based systems are actively researched for their potential to optimize industrial processes. Current approaches often rely on training models with extensive experimental data under specific conditions, followed by deployment, but challenges, like tool wear and machine-specific noise, can limit transferability and necessitate cautious adaptation strategies. This study addresses these issues by proposing a novel method for safe, efficient contextual optimization in industrial settings. The approach balances exploration and exploitation through continual learning, supported by appropriate data management and local approximation techniques. Implemented as a modular software solution for industrial edge control, this method is demonstrated on a steel straightening machine, showcasing its ability to adapt reliably to varying operational environments [12].

While existing DRL-based approaches have achieved some success in image augmentation tasks, their effectiveness for data augmentation in intelligent medical image analysis remains unsatisfactory [13, 14]. To address this, the adaptive sequence-length-based deep reinforcement learning (ASDRL) model for automatic data augmentation (AutoAug) is proposed, introducing a precise reward function that evaluates augmentation transformations more accurately and an intelligent automatic stopping mechanism (ASM) that halts augmentation once optimal performance is

achieved. Extensive experiments on medical image segmentation datasets show that ASDRL-AutoAug significantly outperforms state-of-the-art methods, offering superior performance through accurate reward assessment and adaptive sequence length, demonstrating its potential to enhance medical image analysis [15].

Although RL has achieved remarkable successes in various domains, its application in real-world scenarios is limited due to many methods failing to generalize to unfamiliar conditions. This work addresses the problem of generalizing to new transition dynamics, where the environment's response to the agent's actions changes, such as a robot's mobility being affected by different gravitational forces depending on its mass. Effective generalization requires conditioning an agent's actions on extrinsic state information and contextual information that reflects environmental responses. Despite the recognized need for context-sensitive policies, the architectural integration of context information remains underexplored. This work investigates how context information should be incorporated into behavior learning to enhance generalization. A neural network architecture, the decision adapter, is introduced to generate the weights of an adapter module, conditioning the agent's behavior on context information. The decision adapter extends a previously proposed architecture and demonstrates superior generalization performance across multiple environments. Furthermore, it shows increased robustness to irrelevant distractor variables compared to alternative methods [16].

By integrating contextual factors into reward calculations, the mechanism improves the relevance of retrieved images [17]. This proactive adjustment helps in delivering content that is not only visually appealing, but also contextually appropriate, thus enhancing user satisfaction and engagement with the retrieval system. Contextual adaptation fosters adaptive learning within the agent. It enables the system to continuously refine its retrieval strategies based on real-time contextual cues, improving its ability to anticipate and respond to varying retrieval demands effectively. The ability to adapt rewards based on contextual factors contributes to optimized performance across diverse retrieval tasks and user scenarios. This adaptive capability ensures that the agent can maintain high standards of performance and relevance, even in dynamic and unpredictable environments.

In practice, contextual adaptation operates alongside user feedback within the deep deterministic policy gradients (DDPG) framework. The mechanism leverages both user interactions and contextual cues to dynamically adjust rewards during the agent's training and decision-making processes. This integrated approach not only enhances the system's adaptability, but also supports its ability to deliver tailored and contextually relevant image retrieval solutions.

2.4 Enhanced learning and adaptability

In recent years, researchers have increasingly explored the application of RL algorithms as integral components in addressing various natural language processing (NLP) tasks. Particularly noteworthy is their integration into conversational systems,

leveraging deep neural networks. Researchers conducted a comprehensive review of the current state-of-the-art RL methods within the realm of NLP, with a primary focus on conversational systems, given their growing importance. They provided in-depth descriptions of these NLP challenges and discussed why RL represents a suitable approach for tackling them effectively. In addition, they critically examine the advantages and limitations associated with RL methods in this context [18].

Recent advancements in multimedia streaming applications (MAS) have improved video transmission speed but still face issues like slow access, delays, and inefficiencies. Traditional manual annotation for content retrieval is inaccurate over large databases, leading to a shift toward automatic annotation. This study introduces an automated model that retrieves visually similar images from streams using multi-modal active learning (MAL) combined with a convolutional recurrent neural network (CRNN) to annotate based on features like edges, color, and texture. A deep reinforcement learning (DRL) algorithm validates features, enhancing retrieval performance. Simulation results show that the MAL-DRL model outperforms conventional methods across metrics like retrieval accuracy, sensitivity, specificity, and MAPE [19].

A deep reinforcement learning (DRL) approach improves object detection in low-quality images by enhancing image quality through a reward-driven method rather than extensive retraining. Using an image enhancement tool chain (IETC) and a dueling deep Q network-based tool selector (DDQN-TS) with a 'pass' option, the system adapts to variable image quality conditions. A 'thresh' parameter addresses negative sample inconsistencies, demonstrating effectiveness in challenging settings with noise, fog, and uneven lighting [20].

In summary, we proposed DDPG-SARM (deep deterministic policy gradient with self-adaptive reward mechanism) to address the shortcomings of existing methods in dynamic environments, where user preferences and dataset characteristics are constantly evolving. Traditional methods often struggle to adapt to changing conditions, relying heavily on predefined rules or static reward functions. This limits their ability to improve over time or react to real-time user feedback. In contrast, DDPG-SARM dynamically adjusts rewards based on continuous user interaction, enabling faster adaptation to new data patterns and improving retrieval accuracy and efficiency. By integrating both user feedback and contextual cues into the learning process, DDPG-SARM ensures the agent remains responsive and effective, providing consistent, high-quality retrieval outcomes even in unpredictable environments. This adaptability overcomes the limitations of conventional approaches, which often fail to maintain optimal performance in such dynamic settings.

3 Problem formulation

In this research, we aim to develop a robust image retrieval system for medical images, particularly focusing on the *MedMNIST* datasets. The goal is to leverage *reinforcement learning (RL)* with a *deep deterministic policy gradient (DDPG)* framework, which is coupled with a *self-adaptive reward mechanism (SARM)*. The

system will learn to improve the retrieval of relevant medical images by interacting with a dynamic environment. Formally, we define the following components:

3.1 State space (S)

Let $s_t \in S$ be the state at time t , representing the current query image or a feature vector of an image extracted using a convolutional neural network (CNN).

$$s_t = \text{CNN}(x_t) \quad (1)$$

where x_t in X is the input image at time t .

3.2 Action space (A)

The action $a_t \in A$ represents the selection of an image from the dataset to be retrieved, based on its similarity to the query image.

$$a_t = \pi(s_t) \quad (2)$$

π is the policy network that selects an action given the state.

3.3 Reward function (R)

The reward function $R(s_t, a_t)$ provides feedback on the retrieval accuracy. A *self-adaptive reward mechanism (SARM)* is implemented to adjust the rewards dynamically based on the system's performance and user feedback:

$$R(s_t, a_t) = \lambda_1 \cdot \text{user feedback} + \lambda_2 \cdot \text{system accuracy} + \lambda_3 \cdot \text{contextual factors} \quad (3)$$

where $\lambda_1, \lambda_2, \lambda_3$ are adaptive weighting factors.

3.4 Objective function

The primary objective is to maximize the expected cumulative reward $J(\theta)$, where θ represents the parameters of the policy (actor) network:

$$J(\theta) = \mathbb{E}_{s \sim p_\pi} \left[\sum_{t=0}^T \gamma^t R(s_t, a_t) \right] \quad (4)$$

where γ is the discount factor, p_π is the distribution of states under policy π , and T is the time horizon. To ensure that this objective aligns with the evaluation metrics in the experimental section—specifically *accuracy*, *precision*, *recall*, and *F1-score*—we describe below how each metric connects to our objective function and how these metrics will be measured.

- **Accuracy** Accuracy reflects the correct retrieval rate of relevant images among all retrievals. In terms of our objective function, optimizing the cumulative reward $J(\theta)$ encourages correct action selections, resulting in high accuracy in the retrieval of relevant images. During experimental evaluation, accuracy will be calculated as the proportion of correctly retrieved relevant images out of all retrievals.
- **Precision** Precision measures the relevancy of retrieved images, indicating the proportion of relevant images among those retrieved by the model. Within the objective function, maximizing $J(\theta)$ indirectly promotes high precision by rewarding actions that prioritize the retrieval of relevant images based on similarity to the query image. In experiments, precision will be measured as the ratio of true positive retrievals to the sum of true positives and false positives.
- **Recall** Recall assesses the model's capability to retrieve all relevant images, providing a measure of completeness. In our objective, a higher cumulative reward $J(\theta)$ should correlate with the model's ability to consistently retrieve all relevant images for a given query. Experimentally, recall will be calculated as the ratio of true positive retrievals to the total number of relevant images in the dataset.
- **F1-Score** The F1-score offers a balanced metric that considers both precision and recall. By optimizing the cumulative reward $J(\theta)$, we aim to achieve a balanced retrieval performance that does not overemphasize either precision or recall, thereby achieving a high F1-score. This score will be evaluated as the harmonic mean of precision and recall in our experiments.

Through these measures, the objective function $J(\theta)$ will be evaluated based on its impact on accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the retrieval system's effectiveness. This alignment ensures that the objectives of current study are clearly articulated and can be effectively validated through the experimental results.

3.5 Task characteristics

Data imbalance The dataset exhibits imbalance across different organ classes, affecting model performance. To address this, we apply the synthetic minority oversampling technique (SMOTE) to ensure a more balanced distribution during training:

$$N_c = N_{\max}, \quad \forall c \in \text{Classes} \quad (5)$$

where N_c is the number of samples in class c , and N_{\max} is the size of the largest class after balancing.

3.6 Action-value estimate (Q-function) and critic updates

The Q-function $Q(s_t, a_t)$ is trained to minimize the temporal difference (TD) error, representing the objective for updating the critic network. The TD error is minimized as follows:

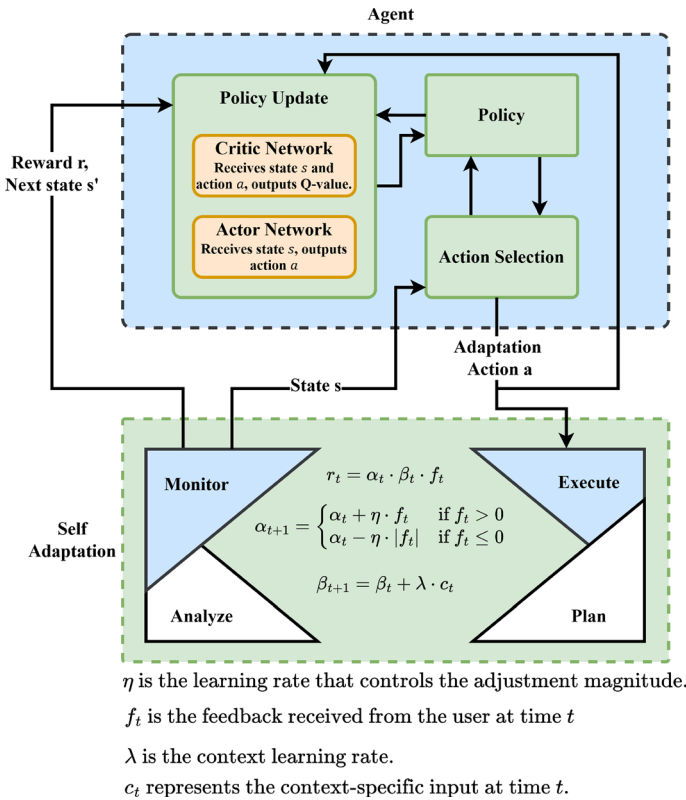


Fig. 2 Proposed framework (DDPG-SARM)

$$\mathcal{L}(\phi) = \frac{1}{N} \sum_{i=1}^N (Q_\phi(s_i, a_i) - (r_i + \gamma Q_\phi(s_{i+1}, a_{i+1})))^2 \tag{6}$$

where ϕ are the parameters of the critic network.

4 Proposed methodology

Figure 2 shows that the self-adaptive reward mechanism dynamically adjusts the rewards provided to the DDPG agent throughout its training process. Instead of relying on static reward schemes that offer fixed rewards based on predefined criteria, this mechanism continuously updates reward values in real-time, utilizing user feedback and contextual cues from the image retrieval tasks. The key advantage of this dynamic reward adjustment lies in its ability to optimize the learning process by promoting adaptability and personalized responses.

The primary objective of this research is to develop a robust image retrieval system utilizing RL. Our approach leverages a convolutional neural network (CNN) for

feature extraction, and a deep deterministic policy gradient (DDPG) agent to learn and optimize the retrieval process. Algorithm 1 presents the overall working of proposed model.

Algorithm 1 DDPG Model with SARM

Inputs: Dataset (training, validation, test splits); batch size; state and action dimensions; maximum action value; training hyperparameters; user feedback; context information.

Outputs: Trained Actor and Critic networks with adaptively modified rewards.

- 1: **Data Transformation and Loading**
- 2: Import libraries and define transformations (resize, normalize, convert to tensor)
- 3: Load dataset splits, apply transformations, and create DataLoaders
- 4: **CNN Model and Feature Extraction**
- 5: Initialize CNN layers (conv1, conv2, conv3, pooling, fully connected layers)
- 6: Extract features by passing dataset through CNN
- 7: **Actor Network Initialization**
- 8: Define layers for Actor (state \rightarrow 400, 400 \rightarrow 300, 300 \rightarrow action)
- 9: Forward pass: apply ReLU activations, then `tanh` on output layer
- 10: **Critic Network Initialization**
- 11: Define layers for Critic (state + action \rightarrow 400, 400 \rightarrow 300, 300 \rightarrow 1)
- 12: Forward pass: concatenate state and action, apply ReLU, output Q-value
- 13: **Initialize SARM**
- 14: Set dynamic adjustment factor and context factor to 1.0
- 15: Define learning rate for updates
- 16: **DDPG Training Loop with SARM**
- 17: **for** each episode **do**
- 18: **for** each time step **do**
- 19: Use Actor to select action based on current state
- 20: Execute action, observe reward, obtain feedback and context
- 21: **Compute Adaptive Reward (SARM)**
- 22: base_reward = user_feedback \times context_factor
- 23: adaptive_reward = base_reward \times dynamic_adjustment_factor
- 24: Store (state, action, adaptive reward, next state) in buffer
- 25: **Update SARM Factors**
- 26: **if** feedback $>$ 0 **then**
- 27: Increase dynamic adjustment factor
- 28: **else**
- 29: Decrease dynamic adjustment factor
- 30: **end if**
- 31: Update context factor based on context
- 32: **Train Actor-Critic Networks**
- 33: Sample mini-batch from buffer
- 34: Update Critic to minimize MSE loss
- 35: Update Actor to maximize Q-values
- 36: **end for**
- 37: Periodically update target networks for Actor and Critic
- 38: **end for**

Deep deterministic policy gradient (DDPG) is a model-free off-policy actor–critic that uses deep function approximators. It is particularly effective in continuous action spaces.

1. Policy update

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta) \quad (7)$$

where α is the learning rate and $J(\theta)$ is the objective function defined as:

$$J(\theta) = \mathbb{E}_{s \sim \mathcal{D}}[Q(s, \pi_{\theta}(s))] \quad (8)$$

2. Q-function update

$$\phi \leftarrow \phi + \beta \nabla_{\phi} (Q_{\phi}(s, a) - y)^2 \quad (9)$$

where β is the learning rate and the target value y is given by:

$$y = r + \gamma Q_{\phi'}(s', \pi_{\theta'}(s')) \quad (10)$$

Here, γ is the discount factor, and ϕ' and θ' are the parameters of the target networks, r is the reward, and s' is the next state.

4.1 Self-adaptive reward mechanism

The self-adaptive reward mechanism shown in Fig. 3 dynamically adjusts the rewards during the learning process based on user feedback and contextual factors, ensuring the DDPG agent's adaptability to evolving environments. This mechanism is particularly effective in image retrieval tasks, where user preferences and dataset characteristics change over time. Dynamic adjustment is controlled by factors that are updated after each interaction, based on user feedback and contextual cues.

The process of adjusting rewards is illustrated in Algorithm 2, outlining the step-by-step computation and updating of adaptive rewards. The reward mechanism initializes two key factors: the dynamic adjustment factor and the context factor, both of which influence how rewards are calculated and adjusted throughout the learning process.

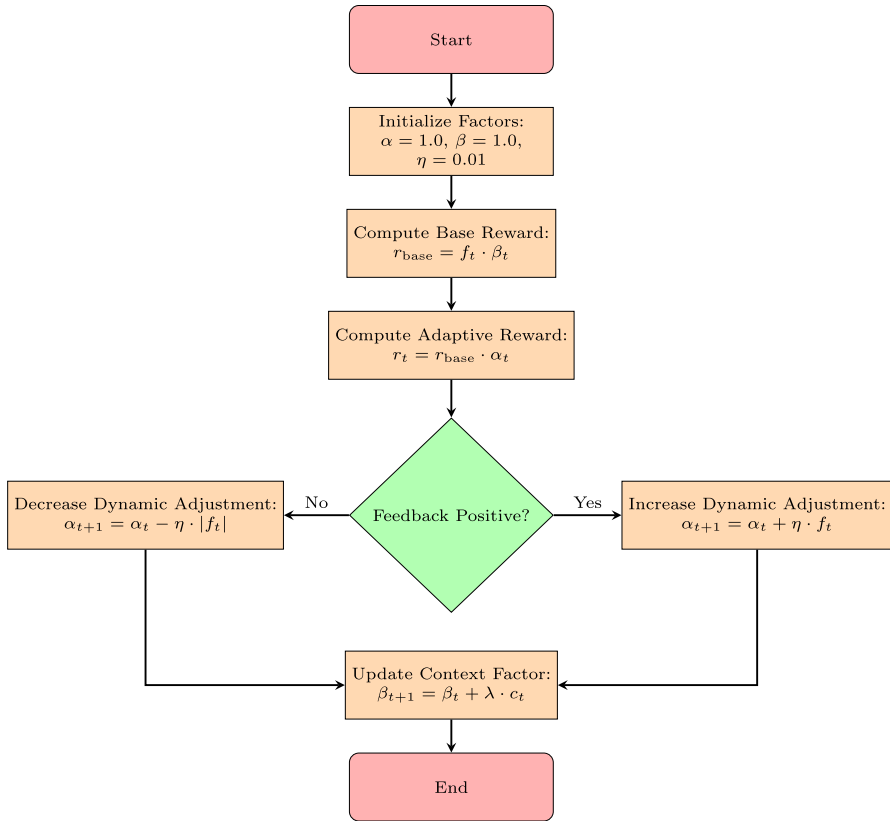


Fig. 3 Self-adaptive reward flowchart

Algorithm 2 Self-adaptive reward mechanism

Inputs: User feedback, Context information

Outputs: Adaptive reward, Updated adjustment factors

- 1: **Initialize** adjustment factor = 1.0, context factor = 1.0, learning rate = 0.01
- 2: **Compute Adaptive Reward:**
- 3: Calculate base reward = feedback × context factor
- 4: Calculate adaptive reward = base reward × adjustment factor
- 5: Return adaptive reward
- 6: **Update Adjustment Factors:**
- 7: **if** feedback is positive **then**
- 8: Increase adjustment factor by (learning rate × feedback)
- 9: **else**
- 10: Decrease adjustment factor by (learning rate × absolute value of feedback)
- 11: **end if**
- 12: Update context factor = context × learning rate

Let r_t be the reward at time step t , f_t be the user feedback, and c_t be the context at that time step. The adaptive reward r_t is computed as:

$$r_t = \alpha_t \cdot \beta_t \cdot f_t \quad (11)$$

where α_t is the dynamic adjustment factor, updated based on user feedback. β_t is the context factor that accounts for contextual information.

4.1.1 Dynamic adjustment factor

The dynamic adjustment factor α_t is updated according to the feedback f_t received. The update rule is:

$$\alpha_{t+1} = \begin{cases} \alpha_t + \eta \cdot f_t & \text{if } f_t > 0 \\ \alpha_t - \eta \cdot |f_t| & \text{if } f_t \leq 0 \end{cases} \quad (12)$$

where η is the learning rate that controls the adjustment magnitude. f_t is the feedback received from the user at time t .

4.1.2 Context factor

The context factor β_t is updated based on the contextual information c_t as follows:

$$\beta_{t+1} = \beta_t + \lambda \cdot c_t \quad (13)$$

where λ is the context learning rate. c_t represents the context-specific input at time t .

In proposed mechanism, user feedback influences the base reward, which is further adjusted by multiplying it with the dynamic adjustment factor. The context factor, which captures contextual information about the task, is also incorporated into the reward calculation. After each interaction, the factors are updated; positive feedback increases the dynamic adjustment factor, encouraging the agent to continue pursuing successful strategies, while negative feedback decreases it, signaling the need to explore alternative strategies. Contextual information is used to fine-tune the context factor, ensuring that the reward mechanism remains sensitive to the evolving environment.

5 Experimental setup

The experimental setup includes details on the dataset(s) used, hyperparameters, evaluation metrics, and the overall process used to train and evaluate the reinforcement learning (RL) based image retrieval system.

5.1 Dataset(s) description

Following subsections discuss the datasets used in current research work.



Fig. 4 Sample images from DermaMNIST dataset

5.1.1 DermaMNIST

The DermaMNIST is based on the HAM10000 [21, 22], which is a large collection of multi-source dermoscopic images of common pigmented skin lesions. The DermaMNIST dataset provides a comprehensive collection of 10,015 images classified into seven different categories, which represent various skin lesions and conditions. Figure 4 presents sample of each category.

Actinic keratoses (Label 0) are rough, scaly patches on the skin that develop from years of exposure to the sun. This category includes both precancerous and cancerous lesions, making accurate classification crucial for early treatment. Basal cell carcinoma (Label 1) is a common type of skin cancer that originates in the basal cells, which are found in the lower part of the epidermis. Benign keratosis-like lesions (Label 2) are noncancerous growths that resemble keratosis, a condition marked by rough, raised skin. Dermatofibroma (Label 3) are benign skin nodules commonly found on the lower legs. Melanoma (Label 4) is a serious and potentially fatal form of skin cancer that arises from melanocytes. Melanocytic nevi (Label 5) has the

Table 1 Class distribution in DermaMNIST dataset

Class	Disease	Dataset		
		Training	Validation	Testing
0	Actinic keratoses	228	33	66
1	Basal cell carcinoma	359	52	103
2	Benign keratosis-like lesions	769	110	220
3	Dermatofibroma	80	12	23
4	Melanoma	779	111	223
5	Melanocytic nevi	4693	671	1341
6	Vascular lesions	99	14	29
Total instances		7007	1003	2005

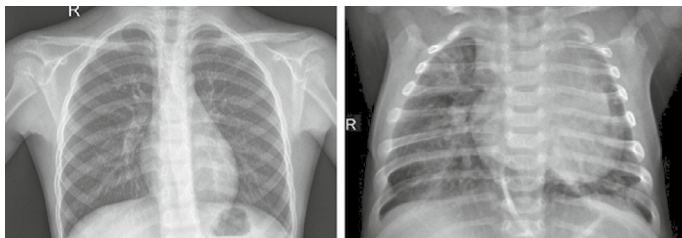


Fig. 5 Sample images from PneumoniaMNIST dataset

highest number of instances, representing benign moles, commonly known as nevi. Vascular lesions (Label 6) are abnormal clusters of blood vessels on the skin and are generally benign.

This distribution in Table 1 shows a significant class imbalance, with melanocytic nevi being the most prevalent and vascular lesions and dermatofibroma being the least common. The presence of both benign and malignant categories in this dataset highlights its value in training models for early and accurate detection of skin conditions, emphasizing the importance of handling class imbalance in model training to ensure robust performance across all classes.

5.1.2 PneumoniaMNIST

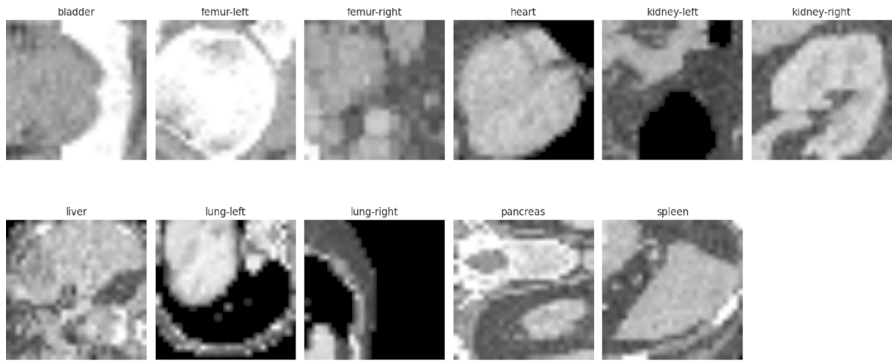
The PneumoniaMNIST is based on a prior dataset [23] of 5856 pediatric chest X-ray images. The PneumoniaMNIST dataset consists of pediatric chest X-ray images and is designed as a binary classification task for detecting pneumonia, with two categories as shown in Fig. 5, representing healthy and diseased lungs.

Table 2 presents the data distribution of PneumoniaMNIST dataset. Normal (Label 0) class includes images of normal, healthy lungs without signs of infection. Identifying normal cases accurately is essential to prevent unnecessary interventions and to ensure that healthy individuals are not misdiagnosed with pneumonia. Pneumonia (Label 1) represents lungs showing signs of infection, such as inflammation caused by bacteria or viruses. With the higher number of instances in this category, the dataset emphasizes pneumonia detection, which is particularly crucial for early treatment and better health outcomes, especially in vulnerable pediatric patients.

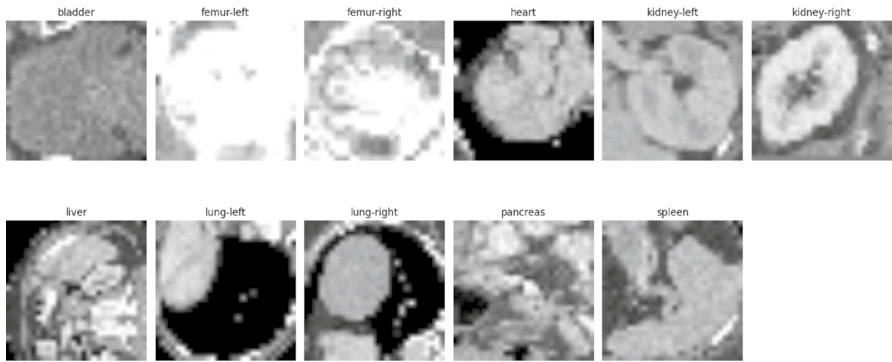
The dataset is highly imbalanced, with more than twice the number of pneumonia instances compared to normal instances. This imbalance is common in medical

Table 2 Class distribution in PneumoniaMNIST dataset

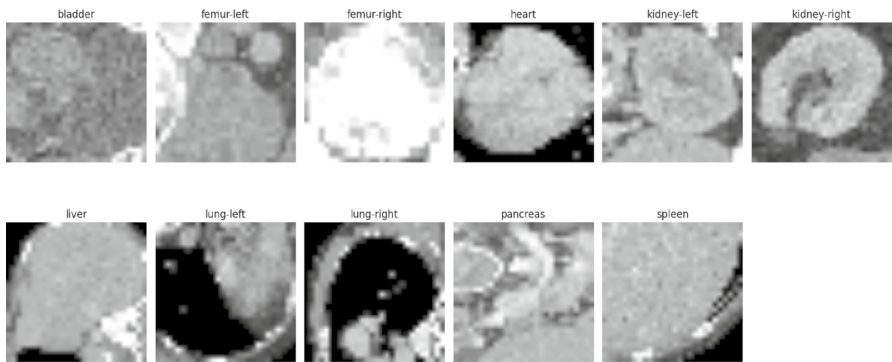
Class	Disease	Dataset		
		Training	Validation	Testing
0	Normal	1214	135	234
1	Pneumonia	3494	389	390
Total instances		4708	524	624



(a) Training samples



(b) Validation samples



(c) Testing samples

Fig. 6 Sample images from OrganMNIST dataset

Table 3 Class distribution in OrganMNIST dataset

Class	Organ	Dataset		
		Training	Validation	Testing
0	Bladder	1956	321	1036
1	Femur-left	1390	233	784
2	Femur-right	1357	225	793
3	Heart	1474	392	785
4	Kidney-left	3963	568	2064
5	Kidney-right	3817	637	1965
6	Liver	6164	1033	3285
7	Lung-left	3919	1033	1747
8	Lung-right	3929	1009	1813
9	Pancreas	3031	529	1622
10	Spleen	3561	511	1884
Total instances		34,561	6491	17,778

datasets, where diseased cases are often prioritized for detection. Effective handling of this imbalance is critical for developing a reliable model that can accurately distinguish between normal and pneumonia cases, helping to reduce false negatives and ensure appropriate care for those affected.

5.1.3 OrganMNIST

The OrganMNIST is based on 3D computed tomography (CT) images from liver tumor segmentation benchmark (LiTS) [24] from MedMNIST v1 [25]. Hounsfield-unit (HU) of the 3D images is transformed into grayscale with an abdominal window [26]. Figure 6 presents the samples from OrganMNIST dataset.

Table 4 Hyperparameters and their settings

Learning rates	
Actor network	$\alpha_{\text{actor}} = 3 \times 10^{-4}$
Critic network	$\alpha_{\text{critic}} = 3 \times 10^{-4}$
Batch size	$B = 32$
Discount factor	$\gamma = 0.99$
Soft update parameter	$\tau = 0.005$
Replay buffer size	$N = 100,000$
Max action	$\text{Max_action} = 1.0$
Number of epochs	Epochs = 100
Iterations per training call	Iterations = 100
Dynamic adjustment learning rate (for self-adaptive reward mechanism)	$\eta = 0.01$

Table 3 shows that the dataset contains a total of 58,830 grayscale images, each sized 28×28 pixels, distributed across 11 classes representing different organs. The data set is divided into three parts: training, validation, and test sets. Total Records: 58,830 images, Training dataset: 34,561 images, Validation dataset: 6491 images, Testing dataset: 17,778 images.

The dataset is not balanced. The number of instances per class varies significantly, with the ‘liver’ class having the highest number of instances (6164) and the ‘femur-right’ class having the fewest (1357). This imbalance can affect the performance of classification models, as models may become biased toward the more frequent classes. So we applied the SMOTE [27] to balance the data, after sampling we get 67,804 each class consists of 6164 instances

5.2 Hyperparameters

Hyperparameters play a critical role in the performance of both the feature extractor (CNN) and the RL agent (DDPG). Table 4 presents the key hyperparameters used in our experimental setup:

5.3 Data transformation and loading

The data transformation and loading phase is crucial for preparing the dataset for training. We utilize data augmentation and normalization techniques to ensure that the model generalizes well to unseen data.

Let X be the set of raw images, T be the set of transformations, and X' be the transformed images. The transformation can be expressed as:

$$X' = T(X) \quad (14)$$

where T includes operations like grayscale conversion, resizing, and tensor conversion.

5.4 Neural network model and features extraction

A convolutional neural network (CNN) is used to extract features from the images. This network consists of convolutional layers followed by pooling layers, which help in capturing spatial hierarchies in images.

1. Convolutional layer

$$y = f(W * x + b) \quad (15)$$

where W is the weight matrix, $*$ denotes the convolution operation, x is the input, b is the bias, and f is the activation function (ReLU in this case).

2. Pooling layer

$$y = \max(x) \quad (16)$$

where max denotes the max-pooling operation applied to the input x .

3. Fully connected layer

$$y = f(Wx + b) \quad (17)$$

where W and b are the weights and biases of the layer and f is the activation function (ReLU).

The feature extraction function uses the CNN to transform each image into a feature vector. These feature vectors are then used as inputs to the RL agent. Let CNN be the feature extractor and x be an input image. The feature vector ϕ is given by:

$$\phi = \text{CNN}(x) \quad (18)$$

5.5 Actor and critic networks

The actor–critic method is used in RL where the actor network decides the actions to be taken, and the critic network evaluates the action by computing a value function. In deep deterministic policy gradient (DDPG), the critic loss and actor loss typically exhibit the following behavior:

The critic network is trained to minimize the difference between its Q value estimates and the target Q values, computed using the Bellman equation. Since this is a regression problem where the loss function is the mean squared error (MSE), the critic loss is generally positive. It measures the squared difference between the predicted Q values and the target Q values.

The critic loss is calculated using the mean squared error (MSE) between the predicted Q values and the target Q values:

$$\text{Critic Loss} = \frac{1}{N} \sum_{i=1}^N (Q(s_i, a_i) - (r_i + \gamma Q'(s_{i+1}, \pi'(s_{i+1}))))^2$$

where Q is the critic network, Q' is the target critic network, π' is the target actor network, r_i is the reward, γ is the discount factor, s_i and a_i are the states and actions in the replay buffer.

The actor network is trained to maximize the Q value predicted by the critic for the actions it selects. In practice, this means minimizing the negative of the expected Q value. Therefore, the actor loss is often negative, indicating that the policy's actions are achieving higher Q values as predicted by the critic.

The actor loss is calculated as the negative expected Q value:

$$\text{Actor Loss} = -\frac{1}{N} \sum_{i=1}^N Q(s_i, \pi(s_i))$$

where π is the actor network.

The actor network is updated by minimizing this loss, which means maximizing the Q value predicted by the critic. Since we minimize the negative Q value, the actor loss often appears negative. A negative actor loss means that the Q values for the actions chosen by the actor are positive and potentially increasing.

5.6 Training and validation loop

The training loop involves iterative updating of the actor and critic networks using batches of data from the replay buffer. Validation is performed to assess the model’s performance on unseen data. Figure 6 presents the sample images from the dataset.

1. Training iteration

for each iteration:

$$\begin{aligned}
 & s, a, r, s', d \sim \text{ReplayBuffer} \\
 & y = r + \gamma Q_{\phi'}(s', \pi_{\theta'}(s')) \\
 \text{Update critic: } & \phi \leftarrow \phi + \beta \nabla_{\phi} (Q_{\phi}(s, a) - y)^2 \\
 \text{Update actor: } & \theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)
 \end{aligned}$$

2. Validation

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total predictions}} \tag{19}$$

5.7 Evaluation metrics

To evaluate the performance of the image retrieval system, several metrics are considered:

- **Accuracy:**

$$\text{Accuracy} = \frac{\text{Number of correctly retrieved images}}{\text{total number of images}} \tag{20}$$

- **Precision, recall, and F1-score:** these metrics can provide additional insights into the performance of the retrieval system, especially in cases of imbalanced datasets. They are defined as:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{false positives}} \tag{21}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{false negatives}} \tag{22}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

This experimental setup integrates CNN-based feature extraction with a RL approach (DDPG) and a self-adaptive reward mechanism to enhance image retrieval accuracy. The setup is designed to ensure robust training, effective feature extraction, and

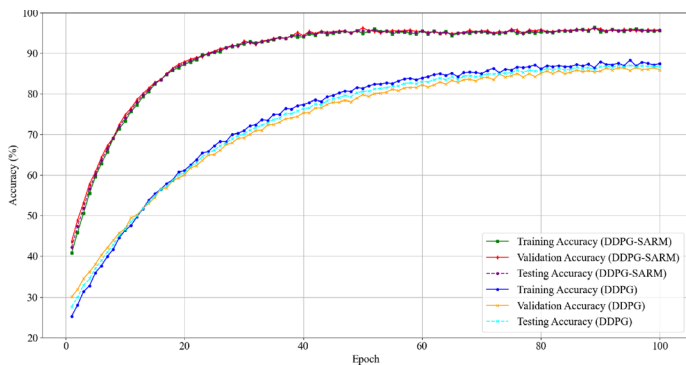


Fig. 7 Training, validation and test accuracies over epochs

adaptive learning based on user feedback. Evaluation metrics like accuracy, precision, recall, and F1-score are used to assess the system's performance comprehensively.

6 Results and discussion

Figure 7 illustrates the training, validation, and test accuracies of DDPG-SARM. Analyzing the trend of these accuracies provides significant insights into the model's learning behavior, generalization capacity, and potential areas for further improvement.

In the initial epochs, both training and validation accuracies rise sharply. The training accuracy starts at approximately 40% and quickly climbs to around 70% by epoch 20. Similarly, the validation accuracy begins slightly higher and reaches around 85% by the same epoch. This rapid increase suggests that the model is effectively learning the fundamental patterns in the data. The fact that the validation accuracy is higher than the training accuracy during this phase might indicate that the validation set contains slightly easier examples or that the model is generalizing well from the onset. The training accuracy gradually rises from 70 to about 80%, while the validation accuracy continues its upward trend but starts to plateau around 90% after epoch 40. This deceleration indicates that the model is entering a phase of fine-tuning where it makes smaller adjustments to optimize performance. In the later stages of training, the training accuracy continues its gradual ascent, eventually surpassing 90% by epoch 100. The validation accuracy, on the other hand, fluctuates slightly around the 90% mark, but remains relatively stable. This convergence between training and validation accuracies is a positive sign, indicating that the model is not overfitting to the training data and is maintaining good generalization to unseen data. The stability in validation accuracy further confirms that the model has effectively learned the dataset and any additional epochs yield marginal improvements.

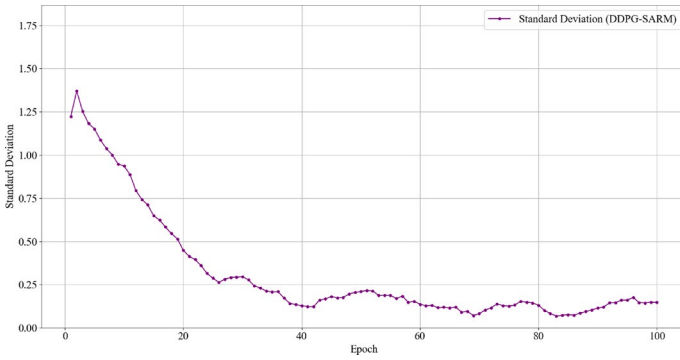


Fig. 8 Standard deviation over epochs

The overall training and validation accuracy trends suggest that the model does not suffer from significant overfitting, as evidenced by the convergence of both accuracies at high values. The model appears to have achieved a good balance between learning the training data and generalizing to the validation set. This balance is crucial for ensuring that the model performs well on new, unseen data. Achieving around 90% accuracy in both training and validation is indicative of strong performance on the task.

Figure 8 shows the standard deviation of accuracy over epochs for the DDPG-SARM model reveals key insights into its learning behavior. Initially, the significant decrease in standard deviation indicates that the model is stabilizing, suggesting an

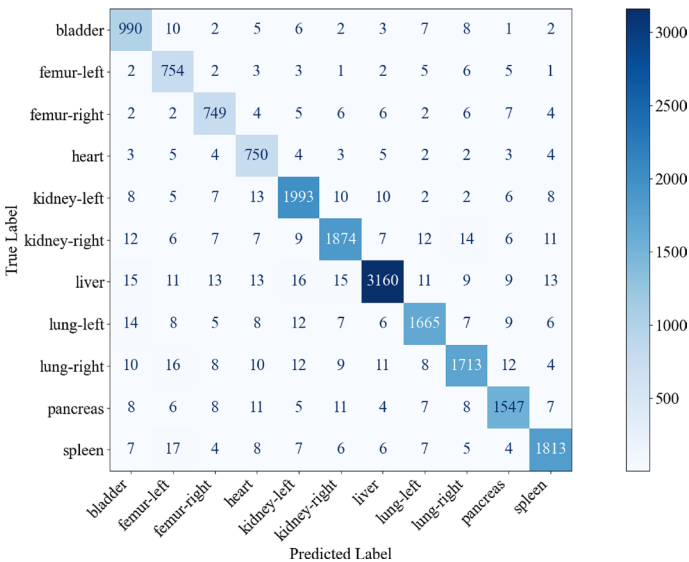


Fig. 9 Confusion matrix on OrganMNIST test dataset

improvement in consistency between training and validation accuracy. This narrowing gap implies enhanced generalization, as the model becomes less sensitive to fluctuations in the data. As the epochs progress, the graph levels off with minimal variability, signaling that the model has reached a stable state where its performance is consistent across both training and validation sets. This stability indicates that the model has been effectively regularized and is less likely to overfit, which is crucial for ensuring robustness and generalizability, allowing it to handle unseen data reliably. Overall, the decreasing standard deviation reflects the model's balanced learning process and adaptability.

The confusion matrix shown in Fig. 9 presents a comprehensive view of the performance of model across different organ classes. For the bladder class, the model has 991 true positives, indicating a high level of accuracy in detecting bladder images. There are minimal misclassifications into other classes, which shows that the model has learned to distinguish bladder images well. The femur-left and femur-right classes show similar performance, with 755 and 752 true positives, respectively. There are a few misclassifications, such as confusion between femur-left and femur-right, which is understandable given their anatomical similarities. However, overall performance remains strong for these classes. In the heart class, there are 746 true positives. There are some misclassifications into classes like bladder and liver, but these are relatively few, indicating that the model performs well in identifying heart images. The kidney-left and kidney-right classes have 1990 and 1877 true positives, respectively. There is slight confusion between these two classes, which can be attributed to the difficulty in distinguishing between the left and right kidneys due to their similar appearances. Despite this, the model still performs well overall in these categories. The liver class stands out with the highest number of true positives

Table 5 Performance comparison with the mainstream models using DermaMNIST, PneumoniaMNIST, and OrganMNIST datasets

Model	DermaMNIST		PneumoniaMNIST		OrganMNIST	
	AUC (%)	OA (%)	AUC (%)	OA (%)	AUC (%)	OA (%)
ResNet-18(28) [28]	91.70	73.50	94.40	85.40	99.70	93.50
ResNet-18(224) [28]	92.00	75.40	95.60	86.40	99.80	<u>95.10</u>
ResNet-50(28) [28]	91.30	73.50	94.80	85.40	99.70	93.50
ResNet-50(224) [28]	91.20	73.10	96.20	88.40	99.80	94.70
Auto-sklearn [29]	90.20	71.90	94.20	85.50	96.30	76.20
AutoKeras [30]	91.50	74.90	94.70	87.80	99.40	90.50
FPViT [31]	92.30	76.60	97.30	89.60	97.60	78.50
EHDfL [32]	91.70	76.90	96.80	88.30	97.40	78.40
CapsNet [33]	94.70	70.72	92.00	84.46	96.35	74.46
ResCaps [34]	95.81	74.56	94.72	88.46	96.13	72.09
3ResCaps [35]	96.10	75.61	95.22	89.42	96.40	74.75
MResCaps [35]	96.25	<u>77.05</u>	96.30	<u>89.74</u>	97.12	79.73
DDPG [36]	96.15	75.45	96.80	89.46	96.80	88.32
Proposed (DDPG-SARM)	<u>96.18</u>	79.65	<u>96.70</u>	94.46	97.85	95.30

Underline and bold values are the highest ones for each metric

at 3169. This indicates excellent performance in detecting liver images, with very few misclassifications into other classes. For the lung-left and lung-right classes, the model achieves 1663 and 1712 true positives, respectively. There are minor misclassifications between these two classes and with kidney-left, but the overall detection performance remains strong. The pancreas class has 1548 true positives, with a few misclassifications. Despite these, the model generally performs well in identifying pancreas images. Finally, the spleen class has 1815 true positives, with minimal misclassifications, indicating robust detection capabilities for spleen images.

6.1 DDPG-SARM versus state-of-the-art (SOTA) methods

Table 5 presents a detailed performance comparison of various models on three medical imaging datasets; DermaMNIST, PneumoniaMNIST, and OrganMNIST. The metrics used for the evaluation include the area under the curve (AUC) and the overall accuracy (OA). The comparison helps to assess how the proposed model, DDPG-SARM, performs against existing models, especially in terms of both classification effectiveness (AUC) and general accuracy (OA) across datasets.

The AUC for DermaMNIST shows the highest values with models using ResCaps variations and reinforcement learning-based models, particularly with CapsNet and MResCaps reaching around 96%. The proposed DDPG-SARM model achieves a high AUC of 96.18%, only slightly behind MResCaps' 96.25%, indicating competitive classification performance. For overall accuracy, the proposed model achieves 79.65%, outperforming all other models. This result suggests that DDPG-SARM improves on general prediction accuracy, likely due to its self-adaptive reinforcement mechanism, which optimizes feature learning dynamically, thus enhancing accuracy.

The AUC values on PneumoniaMNIST are particularly high across models, with FPViT and DDPG reaching the top AUCs of 97.30% and 96.80%, respectively, indicating strong model performance for this task. DDPG-SARM achieves a notable OA of 94.46%, the highest across all models, outperforming the closest model, MResCaps, which has an OA of 89.74%. This significant improvement suggests that the proposed model excels at handling pneumonia classification, possibly due to its advanced adaptive mechanisms that capture critical features more effectively.

OrganMNIST sees high AUC performance across models, with ResNet-18(224) reaching 99.80% AUC, and the proposed DDPG-SARM model performing well with an AUC of 97.85%. The OA metric shows that DDPG-SARM achieves 95.30%, the highest among all models, with ResNet-18(224) following at 95.10%. This improvement in accuracy suggests that DDPG-SARM's performance is consistent and effective across complex image variations within this dataset, likely due to its reinforcement learning framework adapting to subtle feature differences.

The DDPG-SARM model consistently performs at or near the top across all metrics and datasets. It achieves the highest OA on DermaMNIST, PneumoniaMNIST, and OrganMNIST, indicating a strong generalization ability across diverse medical imaging tasks. ResNet-18 and ResNet-50 variants show strong performance in AUC for OrganMNIST but are generally outperformed by DDPG-SARM and some of

the Capsule Network-based models (CapsNet, ResCaps, MResCaps) in DermaMNIST and PneumoniaMNIST. CapsNet, ResCaps, and MResCaps show competitive AUCs, especially for DermaMNIST and PneumoniaMNIST, but their OA values are generally lower compared to DDPG-SARM, highlighting potential overfitting or lesser adaptability to dataset variance. While AutoML models provide a reasonable performance baseline, they do not achieve top performance, especially in OA, which may reflect limitations in AutoML-driven feature adaptation for these medical imaging datasets.

The proposed DDPG-SARM model demonstrates superior performance in overall accuracy across all datasets and provides competitive AUC values, positioning it as a robust and effective model for medical image classification. The reinforcement learning-based approach allows it to dynamically adjust to dataset-specific characteristics, outperforming traditional and AutoML approaches, as well as specialized architectures like capsule networks, especially in terms of accuracy. This makes DDPG-SARM an ideal choice for high-stakes medical applications where both classification performance and reliability are crucial.

6.2 Ablation study

The results obtained from ablation study using OrganMNIST dataset is presented in Table 6 which demonstrates the impact of key components, namely the deep deterministic policy gradient (DDPG) model [36], self-adaptive reward mechanism (SARM), and synthetic minority oversampling technique (SMOTE) [27] on the overall performance of the image retrieval framework. The table compares the system's accuracy, precision, recall, and F1-score across three different configurations.

Configuration 1: DDPG without SARM and SMOTE

In the first row of the table, we observe the performance of the baseline system where DDPG is used without SARM and SMOTE. The results show an accuracy of 85.43%, with precision, recall, and F1-score closely following around 85%. This demonstrates the performance of the base system when the reward mechanism is static and the class imbalance in the data set is not addressed. The relatively lower values across all metrics indicate the limitations of the system when it lacks adaptive learning and a balanced data distribution.

Configuration 2: DDPG with SARM, without SMOTE

When the self-adaptive reward mechanism (SARM) is introduced, we observe an improvement in all performance metrics. The system achieves an accuracy of 88.32%, with precision, recall, and the F1-score also increasing to around 88%. The

Table 6 Ablation study

DDPG	SARM	SMOTE	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
✓	✗	✗	85.43	84.65	85.35	85.11
✓	✓	✗	88.32	87.76	88.65	88.20
✓	✓	✓	95.3	94.76	95.65	95.19

Values in bold text are the highest ones for each metric

inclusion of SARM allows the system to dynamically adjust rewards based on user interactions and context, leading to better decision-making by the DDPG agent. This highlights the importance of SARM in improving adaptability and overall performance, even without addressing data imbalance.

Configuration 3: DDPG with SARM and SMOTE

In the final configuration, both SARM and SMOTE are included, resulting in the highest performance across all metrics. The system achieves an accuracy of 95.3%, with precision at 94.76%, recall at 95.65%, and an F1-score of 95.19%. The introduction of SMOTE effectively addresses the class imbalance in the training dataset, which, combined with the ability of SARM to dynamically adapt rewards, significantly enhances the model's ability to generalize and make accurate predictions. This configuration demonstrates the cumulative benefit of adaptive reward mechanisms and balanced data sampling, which yields optimal performance in the image retrieval task.

The results of the ablation study clearly indicate that both the self-adaptive reward mechanism (SARM) and SMOTE play crucial roles in improving the system performance. The inclusion of SARM alone leads to notable improvements, suggesting that dynamic adjustment of rewards is vital for adapting to diverse user interactions and evolving contexts. Additionally, addressing class imbalance through SMOTE further boosts the system's ability to achieve higher accuracy and balanced prediction outcomes across all metrics.

The highest performing configuration, which combines DDPG with both SARM and SMOTE, highlights the synergy between these components. The dynamic learning capabilities of SARM, along with the balanced data distribution enabled by SMOTE, enable the DDPG agent to better optimize its actions and provide more accurate image retrieval results. The significant jump in performance metrics in this configuration justifies the use of both SARM and SMOTE in the proposed framework.

In summary, the model performs exceptionally well across most organ classes, with high true positive rates and minimal misclassifications. The ability to generalize across different organ types is evident, making the model a reliable tool for medical image retrieval and diagnosis. The few instances of confusion, particularly between anatomically similar organs, suggest areas for further refinement.

7 Conclusion

This research work presents a novel framework for image retrieval using reinforcement learning (RL), specifically employing the deep deterministic policy gradient (DDPG) algorithm enhanced with a self-adaptive reward mechanism. The key findings and contributions of this study underscore the significant advantages of this approach over traditional image retrieval methods and static reward RL techniques. The proposed framework demonstrated superior retrieval accuracy, reaching 95.3%, 94.46%, and 79.65% over OrganMNIST, PneumoniaMNIST, and DermaMNIST datasets, respectively. This improvement is attributed to the dynamic adjustment of rewards, which enables the DDPG agent to learn and adapt effectively to user

preferences and contextual variations. The self-adaptive reward mechanism plays a critical role in this process by continuously updating reward values based on real-time user feedback and contextual cues from retrieval tasks.

Moreover, the framework showed faster convergence, achieving near-optimal performance within 50 epochs. This rapid adaptation is crucial for maintaining high retrieval performance in dynamic environments where user preferences and dataset characteristics evolve unpredictably. In terms of efficiency, the proposed framework exhibited shorter training times due to its adaptive nature, which streamlines the learning process and reduces computational overhead. This efficiency, combined with high accuracy and rapid convergence, highlights the practical applicability of the proposed framework for real-world image retrieval systems. The high precision, recall, and F1-score of the model reflect its effectiveness and reliability in the image retrieval task. These metrics indicate that the model can accurately and comprehensively retrieve relevant images from the dataset, making it well suited for practical applications, particularly in fields where accurate and complete retrieval is critical, such as medical imaging. Further fine-tuning and optimization can enhance these metrics, but current performance suggests a strong and effective model.

8 Future works

Despite the high performance of our current approach, there are always opportunities for improvement. Fine-tuning hyperparameters, such as learning rate, batch size, and network architecture, might yield slight gains in accuracy. Incorporating more sophisticated data augmentation techniques could enhance the model's robustness and generalization further. Additionally, experimenting with regularization methods like dropout or L2 regularization might help mitigate any minor overfitting that is not apparent from the current results. Furthermore, exploring more advanced RL algorithms like soft actor-critic (SAC) and twin-delayed DDPG (TD3) could provide additional stability and efficiency in the learning process. These algorithms have been shown to offer improved learning in complex and high-dimensional environments, which could enhance the effectiveness of the self-adaptive reward mechanism in handling diverse user feedback and contextual data. Future work could involve integrating SAC and TD3 to further refine the adaptability and performance of the image retrieval framework, particularly in more challenging or dynamic scenarios.

In summary, this research validates the effectiveness of using an RL approach with a self-adaptive reward mechanism for image retrieval. The enhanced learning capabilities and adaptability of the DDPG agent result in more personalized and effective retrieval outcomes, improving user satisfaction and engagement. These findings open up new avenues for future research, such as exploring additional datasets, refining the reward mechanism, and incorporating more contextual factors to further enhance the performance of image retrieval systems.

Data availability Data will be made available on request.

Code availability Code will be made available/public on request.

Declarations

Conflict of 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. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Khamaj A, Ali AM (2024) Adapting user experience with reinforcement learning: personalizing interfaces based on user behavior analysis in real-time. *Alex Eng J* 95:164–173
2. Liang T-C, Chang Y-C, Zhong Z, Bigdeli Y, Ho T-Y, Chakrabarty K, Fair R (2024) Dynamic adaptation using deep reinforcement learning for digital microfluidic biochips. *ACM Trans Design Autom Electron Syst* 29(2):1–24
3. Zhu L, Zhang C, Zhang C, Zhang Z, Nie X, Zhou X, Liu W, Wang X (2019) Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. *Appl Soft Comput* 83:105596
4. Hu M, Zhang J, Matkovic L, Liu T, Yang X (2023) Reinforcement learning in medical image analysis: concepts, applications, challenges, and future directions. *J Appl Clin Med Phys* 24(2):13898
5. Wang X, Wang S, Liang X, Zhao D, Huang J, Xu X, Dai B, Miao Q (2024) Deep reinforcement learning: a survey. *IEEE Trans Neural Netw Learn Syst* 35(4):5064–5078
6. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2019) Continuous control with deep reinforcement learning. *arXiv preprint*
7. Zhao H, Tang W, Yao D (2024) Policy optimization for continuous reinforcement learning. *Adv Neural Inform Process Syst* 36
8. Viswanadhapalli JK, Elumalai VK, Shivram S, Shah S, Mahajan D (2024) Deep reinforcement learning with reward shaping for tracking control and vibration suppression of flexible link manipulator. *Appl Soft Comput* 152:110756
9. Huang C, Wang G, Zhou Z, Zhang R, Lin L (2023) Reward-adaptive reinforcement learning: dynamic policy gradient optimization for bipedal locomotion. *IEEE Trans Pattern Anal Mach Intell* 45(6):7686–7695
10. Xu M, Chen X, She Y, Jin Y, Wang J (2024) Time-varying weights in multi-reward architecture for deep reinforcement learning. *IEEE Trans Emerg Topics Comput Intell*
11. Tang Z, Li T, Wu D, Liu J, Yang Z (2024) A systematic literature review of reinforcement learning-based knowledge graph research. *Expert Syst Appl* 238:121880. <https://doi.org/10.1016/j.eswa.2023.121880>
12. De Blasi S, Bahrami M, Engels E, Geppert A (2024) Safe contextual Bayesian optimization integrated in industrial control for self-learning machines. *J Intell Manuf* 35(2):885–903
13. Xu J, Zhang H, Qiu J (2022) A deep deterministic policy gradient algorithm based on averaged state-action estimation. *Comput Electr Eng* 101:108015

14. Zhang W, Chen Q, Yan J, Zhang S, Xu J (2021) A novel asynchronous deep reinforcement learning model with adaptive early forecasting method and reward incentive mechanism for short-term load forecasting. *Energy* 236:121492. <https://doi.org/10.1016/j.energy.2021.121492>
15. Xu Z, Wang S, Xu G, Liu Y, Yu M, Zhang H, Lukasiewicz T, Gu J (2024) Automatic data augmentation for medical image segmentation using adaptive sequence-length based deep reinforcement learning. *Comput Biol Med* 169:107877
16. Beukman M, Jarvis D, Klein R, James S, Rosman B (2024) Dynamics generalisation in reinforcement learning via adaptive context-aware policies. *Adv Neural Inform Process Syst* 36
17. Yang R, Pan X, Luo F, Qiu S, Zhong H, Yu D, Chen J (2024) Rewards-in-context: multi-objective alignment of foundation models with dynamic preference adjustment. <https://arxiv.org/abs/2402.10207>
18. Uc-Cetina V, Navarro-Guerrero N, Martin-Gonzalez A, Weber C, Wermter S (2023) Survey on reinforcement learning for language processing. *Artif Intell Rev* 56(2):1543–1575
19. Dhiman G, Kumar AV, Nirmalan R, Sujitha S, Srihari K, Yuvaraj N, Arulprakash P, Raja RA (2023) Multi-modal active learning with deep reinforcement learning for target feature extraction in multi-media image processing applications. *Multim Tools Appl* 82(4):5343–5367
20. Ye J, Wu Y, Peng D (2024) Low-quality image object detection based on reinforcement learning adaptive enhancement. *Pattern Recogn Lett* 182:67–75
21. Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 5(1):1–9
22. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, Helba B, Kallou A, Liopyris K, Marchetti M, et al. (2019) Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint*
23. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172(5):1122–1131
24. Bilic P, Christ P, Li HB et al (2023) The liver tumor segmentation benchmark (lits). *Med Image Anal* 84:102680. <https://doi.org/10.1016/j.media.2022.102680>
25. Yang J, Shi R, Ni B (2021) Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp 191–195. <https://doi.org/10.1109/ISBI48211.2021.9434062>
26. Xu X, Zhou F, Liu B, Fu D, Bai X (2019) Efficient multiple organ localization in CT image using 3D region proposal network. *IEEE Trans Med Imaging* 38(8):1885–1898. <https://doi.org/10.1109/TMI.2019.2894854>
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 16(1):321–357
28. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 770–778
29. Feurer M, Klein A, Eggenberger K, Springenberg J, Blum M, Hutter F (2015) Efficient and robust automated machine learning. *Adv Neural Inform Process Syst* 28
30. Jin H, Song Q, Hu X (2019) Auto-keras: an efficient neural architecture search system. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19*, pp 1946–1956. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3292500.3330648>
31. Liu J, Li Y, Cao G, Liu Y, Cao W (2022) Feature pyramid vision transformer for medmnist classification decathlon. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp 1–8. IEEE
32. Han Q, Hou M, Wang H, Wu C, Tian S, Qiu Z, Zhou B (2023) EHDFL: Evolutionary hybrid domain feature learning based on windowed fast Fourier convolution pyramid for medical image classification. *Comput Biol Med* 152:106353
33. Mukhometzianov R, Carrillo J (2018) CapsNet comparative performance evaluation for image classification. *CoRR* [arXiv:abs/1805.11195](https://arxiv.org/abs/1805.11195)
34. Ai X, Zhuang J, Wang Y, Wan P, Fu Y (2022) ResCaps: an improved capsule network and its application in ultrasonic image classification of thyroid papillary carcinoma. *Complex Intell Syst* 8(3):1865–1873
35. Sengul SB, Ozkan IA (2024) MResCaps: enhancing capsule networks with parallel lanes and residual blocks for high-performance medical image classification. *Int J Imaging Syst Technol* 34(4):23108

36. Farooq A, Zhang X (2023) Tongue image retrieval based on reinforcement learning. In: Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition. ICCPR '22, pp 282–289. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3581807.3581848>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Farooq Ahmad¹ · Xinfeng Zhang¹ · Zifang Tang² · Fahad Sabah² · Muhammad Azam³ · Raheem Sarwar⁴

✉ Raheem Sarwar
r.sarwar@mmu.ac.uk

Farooq Ahmad
farooq.ahmad.bjut@hotmail.com

Xinfeng Zhang
xfzhang@bjut.edu.cn

Zifang Tang
zifangtang@emails.bjut.edu.cn

Fahad Sabah
fahad.sabah@hotmail.com

Muhammad Azam
pmit@superior.edu.pk

- ¹ College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
- ² College of Computer Science, Beijing University of Technology, Beijing 100124, China
- ³ Faculty of CS and IT, Superior University, Lahore 54000, Pakistan
- ⁴ OTEHM, Faculty of Business and Law, Manchester Metropolitan University, Manchester M15 6BH, UK