Active Deep Reinforcement Learning to Support Breast Cancer Risk Labeling

Giulia Zanon Castro
Graduate Program in Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
giuliaz@ufmg.br

Cassia Rita Pereira da Veiga
Department of Health Services Management
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
cassia.veig@gmail.com

Frederico Gadelha Guimarães
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
fredericoguimaraes@ufmg.br

Abstract—**Breast cancer is one of the most common cancers in women worldwide. Early detection of breast cancer plays a key role in alleviating pressure on healthcare services, decreasing morbidity and mortality, and minimizing the economic burden on public health systems. While machine learning (ML) has shown promise in medical imaging, its effectiveness is limited by the scarcity of labeled data. In this context, this study explores the use of Reinforcement Learning (RL) to optimize sample selection for annotating mammographic images. Using a subset of the CBIS–DDSM dataset, which is available on Kaggle, we implemented an RL–based strategy combined with an InceptionV3 model pre–trained on RadImageNet for feature extraction. A custom environment was developed in which the RL agent learns to prioritize the most informative samples, aiming to reduce annotation costs in resource–constrained settings. Results show that the RL–based approach outperforms random sampling, achieving comparable accuracy to the fully supervised baseline (71% F–score) using fewer labeled samples. The RL model exhibited a consistent improvement trend without saturation, indicating potential for further gains with continued training. By addressing the challenge of efficient labeling, this work contributes to the development of more generalizable and cost–effective diagnostic models. The approach is particularly relevant in multi–institutional scenarios with limited and heterogeneous data, and can be further enhanced by incorporating additional clinical information to improve diagnostic precision.**

Index Terms—**machine learning, reinforcement learning, active learning, computer vision, data labeling, breast cancer, health care**

I. Introduction

Machine learning (ML) solutions in healthcare have demonstrated promising performance in research [1], [2]. However, translating these models to real clinical settings presents challenges, including data acquisition and annotation. While models often require large amounts of labeled data to maintain their effectiveness, the process of mass annotation in clinical contexts is frequently impractical due to time constraints, cost, and the scarcity of expert annotators. Rather than simply adding more data, which has become increasingly challenging due to the explosive growth in demand for annotations, the emphasis is now on achieving better performance with fewer but more strategically selected data for training [3], [4]. This approach involves selecting a part of the available data for annotation, which can serve as a representative proxy for the entire dataset, saving resources [3]. By addressing the practical constraints of data labeling in clinical environments, this strategy aims to bridge the gap between the performance of ML models in controlled settings and their applicability in real–world healthcare scenarios.

In response to these challenges, various ML approaches have been explored to enhance model adaptability and efficiency in clinical settings. Among these, reinforcement learning (RL) has emerged as particularly advantageous in dynamic clinical contexts [5], such as oncology. RL models can continuously adapt to evolving patient states and integrate expert feedback directly into the learning process. This adaptability is helpful in complex medical domains like breast cancer, which, given its multiple stages and substantial public health impact, represents a clinical area that greatly benefits from such adaptive approaches.

The application of RL in breast cancer detection and management exemplifies its potential in healthcare. Recent studies suggest that RL–based policies can contribute to earlier and more efficient detection of breast cancer [6], directly addressing the need for improved diagnostic accuracy and timeliness. Moreover, RL's adaptability allows for the incorporation of human preferences into the decision–making process. For instance, by penalizing critical diagnostic errors, these models can align more closely with clinical priorities, fostering safer and more context–aware decision–making [7]. This perspective shifts the focus beyond predictive accuracy alone, emphasizing the overall effectiveness and responsibility of AI–driven interventions in healthcare. By integrating expert knowledge and adapting to specific clinical contexts, RL models offer a more reliable approach to medical decision support.

While the application of RL in breast cancer detection demonstrates its potential in specific diagnostic scenarios, its utility in healthcare extends far beyond this single domain. RL's potential spans a wide range of clinical tasks, showing promise in automating and optimizing various aspects of healthcare delivery. From maintaining medical records and interpreting imaging to supporting clinical diagnoses and disease prediction, RL has demonstrated versatility across different layers of the healthcare workflow [8].

This broad applicability is useful in areas such as behavioral

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

health, where RL has been applied to personalize patient communication strategies. For example, RL algorithms have been used to dynamically tailor text messages to promote medication adherence among individuals with diabetes [9]. This application showcases RL's ability to learn from prior interactions and adapt its approach to guide future actions, ultimately enhancing patient engagement and improving outcomes in longitudinal care settings.

In this context, this work proposes an approach that combines Active Learning (AL) with RL to address the challenges of limited labeled data in clinical settings. This integration is relevant in healthcare contexts, where data annotation is costly and time–consuming, and not all samples contribute equally to a model's learning process. By leveraging RL's adaptability and AL's strategic sample selection, this approach aims to optimize the annotation process, making it more efficient and cost–effective. Traditional AL methods typically employ static criteria, such as uncertainty, to decide which examples should be labeled [10]. However, they fail to adjust to how these selections influence the model's performance over time. By integrating RL, it becomes feasible to create an agent capable of learning an optimal selection policy dynamically.

This RL–based policy is guided by rewards based on the enhancement in model performance with each newly added labeled example. Consequently, the agent evaluates not only the present condition of a sample but also the series of prior decisions and their aggregate consequences, leading to a more adaptive and efficient labeling approach. By leveraging RL's ability to learn from interactions and adapt its strategy and combining it with AL's focus on selecting the most informative samples, this method aims to maximize the value of limited annotation resources.

While this study utilizes a dataset with available labels for all samples, these labels are employed only for assessment and simulation purposes, mimicking real–world scenarios where complete labeling is often unfeasible. Instead of relying on manual selection or predetermined heuristics, the task of choosing samples for annotation is entrusted to an RL agent. This approach is designed to enhance the usability and adaptability of ML models across various healthcare settings, different data distributions and clinical practices characterize each.

This dynamic strategy is particularly beneficial in healthcare settings where data characteristics and clinical priorities are constantly evolving, ensuring that the model remains relevant and practical across different hospital environments. By addressing the practical constraints of data labeling in clinical environments, this approach aligns with the broader goal of achieving better model performance with strategically selected training data, as discussed in the context of breast cancer detection and other healthcare applications.

Thus, our main contributions are:

- We propose an approach for intelligent sample selection in label–restricted scenarios, using Reinforcement Learning;
- We evaluate how close it comes to the performance that would be achieved with a full labeled set.

II. Background

Reinforcement Learning (RL) techniques have shown promising applications in various medical fields, including the optimization of diagnostic processes, treatment strategies, and medical image segmentation [11]–[16]. In mammography, RL can be used to balance the early detection of breast cancer with the costs and burden on healthcare services [6]. The model learns policies that maximize timely disease detection without generating unnecessary excess screenings.

Similarly, in oncology, RL has been explored to support diagnoses and assist in defining therapeutic strategies. A recent review highlights the technique's potential at multiple points in the clinical journey [7]. This advancement is directly linked to the evolution of offline RL methods, which are designed for scenarios where direct interaction with the environment is impractical. From retrospective databases, algorithms are trained to reproduce previous behaviors and propose more effective decisions, respecting safety, cost, and time constraints typical of clinical practice [17].

RL can also be adjusted to incorporate human preferences in sensitive diagnoses, as in the case of skin cancer [11]. By using tables with clinical estimates of risk and benefit associated with different diagnostic errors, the model learns to prioritize decisions that minimize critical errors, resulting in greater sensitivity in detecting melanomas.

In critical care environments, such as intensive care units (ICUs), RL has been studied as a tool to support real–time decisions based on large volumes of historical data. A recent systematic review points to the growing presence of these techniques in highly complex contexts, where personalized conduct can directly impact clinical outcomes [5].

To enhance the quality of data used in training machine learning models, active learning has been employed. This approach can be used to improve the performance of generative models by focusing on more informative samples during training [4]. In the medical field, [18] evaluate the use of active learning for tasks that require annotations, employing techniques such as entropy or model loss in the labeling decision process.

Building upon these uncertainty–based approaches, [19] applied specific Active Learning methods, namely BALD and Max Entropy, to the domain of dental imaging. Their work focused on selecting the most informative CBCT image slices for multi–label segmentation of dental structures and lesions, demonstrating the practical application of these techniques in specialized medical contexts.

Active Learning can be applied in three main scenarios, as described by [3]: Membership Query Synthesis, Stream–based Selective Sampling, and Pool–based Active Learning. Each scenario has its own characteristics and applications, with Pool–based AL being the most common configuration.

A challenge in active learning is the cold–starting problem, where no labeled data are initially available. This issue is particularly relevant in medical contexts, where obtaining labeled data can be costly and time–consuming. To address this problem, [20] proposes a method for the initial selection of informative samples. Their approach tackles two main issues: biased query, where active sampling methods tend to select samples from majority classes, resulting in imbalanced initial datasets, and outlier query, where unrepresentative data can be selected due to the lack of a trained classifier.

The solution involves extracting representations from the data using a self–supervised contrastive learning technique, which generates discriminative unlabeled embeddings [20]. Using the K–means clustering algorithm, pseudo–labels are generated from these embeddings, ensuring class diversity in the initial selection. To address outliers, [20] employ a modified dataset map based on these pseudo–labels, which enables the identification and prioritization of data that the model has greater difficulty distinguishing from other examples. These samples capture common patterns in the dataset and are more representative and typical of the general distribution, thus providing a coherent starting point for the AL process.

The labeling problem also arises in the context of federated learning, where local data in different institutions can vary significantly in distribution. To address this, [21] proposes an approach for selecting which unlabeled samples should be annotated locally, maximizing information gain while reducing annotation effort. The method combines federated learning, which trains a global model without centralizing sensitive data, with an active sampling strategy based on evidential uncertainty.

In the field of High–Level Synthesis (HLS), which involves the process of converting high–level code into hardware descriptions, [22] proposes an automated code annotation framework that employs RL to explore various combinations of annotations within the source code. In this framework, RL is used to automatically explore different combinations of annotations, referred to as pragmas. The RL agent inserts a set of pragmas, observes the impact on hardware latency generated by the HLS tool, and receives a reward proportional to the improvement obtained. Through this process, the agent progressively learns which pragmas to apply, where to insert them, and in what quantity. The goal is to maximize the performance of the final hardware, even when the original code was not manually optimized. This application demonstrates the versatility of RL techniques beyond medical contexts, illustrating how they can be applied to optimize complex processes in computer engineering.

In the context of medical text processing, [23] propose a method for named entity recognition (NER) in Chinese that utilizes distant supervision in conjunction with reinforcement learning to address incomplete and noisy annotations. Combined distant supervision consists of using a small manually labeled dataset (H) together with a large unannotated corpus (U), where an entity dictionary (D) is used to identify and automatically label mentions in U, assuming that any string matching an entry in D can be an entity, thus generating an automatically annotated set (A). This approach enables the creation of annotated data without human intervention, but it introduces false positive examples, which are considered noise, as well as false negative examples. To reduce the noise, an RL–based method is employed, where an agent analyzes sentences from the combined set H ∪ A, decides whether to include each automatically annotated sentence and is trained with rewards based on the performance of a NER model on the given instances, making decisions that improve the quality of the supervised training.

RL has also been used to improve clinical reasoning ability and generalization in complex tasks, such as Visual Question Answering (VQA) in medical images [12]. The methodology proposed by [12] optimizes the generation of correct answers and logical structure in VQA tasks using the Group Relative Policy Optimization (GRPO) technique [24]. RL is used to explore different forms of reasoning and select those that result in greater clinical coherence and higher accuracy. At each iteration, the model generates multiple answers to a medical question, receives rewards based on objective criteria, and adjusts its policy to prefer more effective answer styles. This approach enables the model to enhance its performance even with limited data and high semantic complexity, which is particularly useful in the medical domain, where high–quality annotated data can be scarce, and questions often require complex reasoning.

In addition to healthcare, Active Reinforcement Learning with Differential Evolution has been successfully applied in a steganalysis task focused on digital security [10]. The task involves analyzing digital media to investigate the insertion of hidden, visually imperceptible data within these digital media. This approach enables the network to initially focus on samples with high entropy loss and then shift to more complex and underexplored samples as the classifier's accuracy improves, demonstrating the potential of combining active learning and reinforcement learning in diverse domains.

Thus, although RL applications in healthcare are promising, the success of these models depends directly on the representativeness of the training data. Given the high time and financial cost required for medical data annotation, there is a need to shift focus from quantity to prioritizing quality in training datasets. With this idea, we propose a strategy that combines active learning and reinforcement learning techniques for intelligent sample selection. By prioritizing data quality over quantity, our strategy not only enhances the efficiency of machine learning models in healthcare but also contributes to more sustainable machine learning applications in healthcare.

III. Methods

A. Dataset

To validate the active labeling strategy with reinforcement learning, the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset [25] was used, which consists of a subset of the Digital Database for Screening Mammography (DDSM) research on breast cancer. The original DDSM consists of 2,620 mammography studies from digitized film.

To generate this subset, CBIS-DDSM underwent a curation process, which included decompressing the original images, converting them to a more accessible format, adding Region of Interest (ROI) segmentations, and adding bounding boxes. Pathological diagnoses were also included in the training data.

We used a curated subset of the CBIS-DDSM dataset[1], available on Kaggle, which provides a reduced and preprocessed version of the original dataset (over 160 GB), making it more accessible for experimentation and reproducibility.

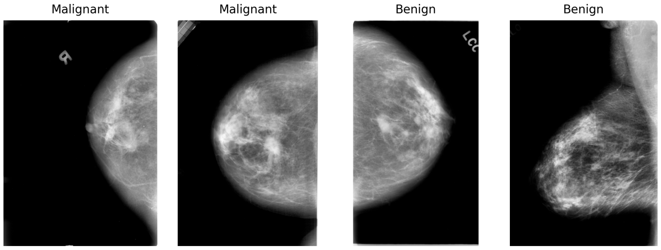

Figure 1. Overview of the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [25] dataset used for validating the active labeling strategy.

B. Baseline Model

This section outlines the methodology for classifying breast cancer into malignant and benign categories using Active Learning (AL) and Reinforcement Learning (RL). First, we develop a baseline model to evaluate whether sample selection techniques could achieve performance comparable to the baseline using a smaller and more informative subset of labeled data.

We first performed a stratified split of the entire dataset into a training + validation set (80%) and a hold-out test set (20%), ensuring that each subset preserved the original class proportions. Next, we applied a stratified split to the 80% portion, allocating 64% of the total data for training and 16% for validation to optimize hyperparameters. All mammogram images were resized, converted from grayscale to RGB, and normalized to the range $[-1, 1]$.

For the baseline, we explored two architectures, DenseNet121 [26] and InceptionV3 [27], which were pre–trained on RadImageNet, a medical image database [28]. RadImageNet is a large–scale dataset that includes annotated ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) images. It contains information on anatomical regions and pathological conditions and is helpful for domain pre–training in medical imaging tasks.

Thus, using the DenseNet121 and InceptionV3 pre–trained in the medical domain, we explored different strategies for freezing and fine–tuning convolutional blocks to balance generalization with adaptation to the specific task of breast cancer risk classification. Most convolutional layers were initially frozen to retain the domain representations learned from RadImageNet, and only deeper layers were gradually unfrozen during training. In all configurations, the original classification head was removed and replaced with a custom architecture. The final hyperparameter settings for each model are detailed in the Results section.

The classification head added to the network consists of a global average pooling layer, followed by the injection of Gaussian noise as a regularization strategy. Subsequently, we interleave dense layers with ReLU activations, batch normalization, and dropout to improve generalization and reduce overfitting. The final output layer employs a softmax activation to generate class pseudo–probabilities.

To further reduce overfitting, we applied an early stopping strategy, halting training if no improvement in validation loss was observed over 50 consecutive epochs. Additionally, we applied L2 regularization [29] with an exponential decay schedule to the dense layers, gradually reducing the penalty strength as training progressed. A learning rate decay strategy was also employed to gradually reduce the learning rate across epochs, enhancing convergence stability.

The final baseline evaluation was performed on the held-out test set, providing a reference performance using the fully labeled dataset. Once the baseline was established, we applied AL and RL strategies to investigate whether comparable performance could be achieved using fewer labeled samples, assessing the feasibility of reducing annotation effort without compromising model effectiveness.

After training the baseline model using the full labeled dataset, we simulated a data scarcity scenario by selecting only a small and representative fraction of the training data to initialize the learning process. This reduced initial set was used to train the first model, which then served as the starting point for subsequent cycles of Active Learning and Reinforcement Learning. In each iteration, additional samples were selected based on their expected informativeness, allowing the model to improve progressively. This approach aimed to assess the extent to which model performance could be preserved or improved through strategic, incremental data selection.

C. Reinforcement Learning

For the Active Learning and Reinforcement Learning experiments, we simulate a scenario of limited annotation resources by initially selecting a small, stratified subset of 5% of the labeled data. This subset consists of images from distinct patients and serves as the starting point for training a supervised model. The goal is to reduce the amount of labeled data required while maintaining or improving classification performance.

To facilitate the training process, we first extracted the image embeddings using the InceptionV3 model pre–trained on RadImageNet. All images were pre–processed to present an input format compatible with the InceptionV3 architecture (299×299 pixels) in RGB color space and with normalization in the interval [-1, 1]. Thus, the embeddings for the breast cancer images were obtained, excluding the classification header, and used as representations of the input data for the RL environment.

To structure the RL–based sample selection process, we implemented a custom environment using the Gym interface [30]. In this environment, each unlabeled sample is represented

[1] https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset

by a state composed of image embeddings. The action space consists of binary decisions: request annotation (1) or skip (0). If a sample is annotated, it is added to the training set along with its ground–truth label, simulating the action of a human expert. Algorithm 1 represents the process.

Algorithm 1: Active Deep Reinforcement Learning

```
Load $D_{\text{labeled}}$, $D_{\text{unlabeled}}$, $D_{\text{validation}}$;
Initialize supervised model $M_{\text{sup}}$ and train on $D_{\text{labeled}}$;
Evaluate $acc_{\text{previous}}$ on $D_{\text{validation}}$;
Define RL environment:
    State: $s$ = embedding(image);
    Action: $a \in \{0 : \text{Ignore}, 1 : \text{Request Label}\}$;
    Reward: $r = \Delta$ accuracy on
      $D_{\text{validation}} + \alpha \times$ prediction entropy;
Initialize DQN agent and replay buffer;
for episode $e = 1$ to $N_{episodes}$ do
    Reset environment; copy datasets for episode;
    Set $acc_{\text{current}} \leftarrow acc_{\text{previous}}$;
    while $D_{unlabeled}$ not empty do
        Get sample $x$, compute state $s$;
        Select action $a \leftarrow Agent.ChooseAction(s)$;
        if $a = 1$ (Request Label) then
            Get label $y$ for $x$; update $D_{\text{labeled}}$; remove $x$
              from $D_{\text{unlabeled}}$;
            Retrain $M_{\text{sup}}$ and evaluate new accuracy;
            $r \leftarrow acc_{\text{new}} - acc_{\text{current}}$;
            $acc_{\text{current}} \leftarrow acc_{\text{new}}$;
        else
            Remove $x$ from $D_{\text{unlabeled}}$;
            $r \leftarrow 0$;
        Get next sample $x_{\text{next}}$, compute $s_{\text{next}}$;
        Store transition $(s, a, r, s_{\text{next}})$ in buffer;
        if buffer has enough samples then
            Train agent using batch from buffer;
    Record $acc_{\text{current}}$;
```

The model is trained iteratively. The agent decides whether to label a sample from the unlabeled set in each cycle based on its current policy. After training the model with the updated labeled set, we evaluate its performance using the validation data. The reward signal provided to the agent is defined as the change in validation accuracy resulting from including the newly labeled sample weighted by the entropy value (Equations 1 and 2). Figure 2 shows the process diagram.

$$r_t = (\text{Acc}_{t+1} - \text{Acc}_t) + \alpha \cdot \mathcal{H}(p_t) \tag{1}$$

$$\mathcal{H}(p_t) = -\sum_{c=1}^{2} p_t^{(c)} \log_2(p_t^{(c)}) \tag{2}$$

Where r_t represents the reward at step t, Acc_t the accuracy of the classifier before labeling the sample at step t, Acc_{t+1} the accuracy after including the new sample, α the hyperparameter that controls the entropy weight, $\mathcal{H}(p_t)$ the entropy of the prediction for the current sample and p_t the vector of probabilities predicted by the classifier for the current sample.

The agent employs a Deep Q–Network (DQN) [31] . This neural network estimates the Q–value function $Q(s, a)$, which represents the expected cumulative reward of taking action a in state s. The DQN is trained to approximate the following value function:

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a\right] \tag{3}$$

Where $\mathbb{E}_{\pi}[\cdot]$ denotes the expected value when following the policy π. The state s is a representation of the current unlabeled sample, composed of its image embedding and associated clinical features; the action a corresponds to the binary decision to request label ($a = 1$) or skip ($a = 0$) the current sample. The reward r_t is defined as the change in validation accuracy of the supervised model after labeling the sample at step t, added to the entropy value of the prediction. Finally, $\gamma \in [0, 1]$ is the discount factor that weighs future rewards.

To derive the agent's behavior, we define the policy π that selects the action with the highest estimated Q–value and maximizes the expected return:

$$\pi(s) = \arg\max_{a} Q(s, a) \tag{4}$$

To balance exploration and exploitation during training, an ϵ-greedy strategy is employed:

$$\pi(a \mid s) = \begin{cases} \arg\max_a Q(s, a), & \text{with probability } 1 - \epsilon \\ \text{random action}, & \text{with probability } \epsilon \end{cases} \tag{5}$$

Q-values are updated using the Bellman equation, with a separate target network for improved stability:

$$y = r + \gamma \max_{a'} Q_{\text{target}}(s', a') \tag{6}$$

The DQN is trained to minimize the mean squared error between the estimated Q-value and its target:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}}\left[(y - Q(s, a; \theta))^2\right] \tag{7}$$

where θ are the parameters of the Q-network, and $\mathcal{D}$ is the replay buffer containing past transitions (s, a, r, s').

After each episode, we reevaluated the model, and the updated reward is used to refine the agent's selection policy. Over time, the agent learns to prioritize labeling the most informative samples that yield the highest performance gain, thereby minimizing the annotation cost. This iterative process continues until the labeling budget is exhausted or performance converges.

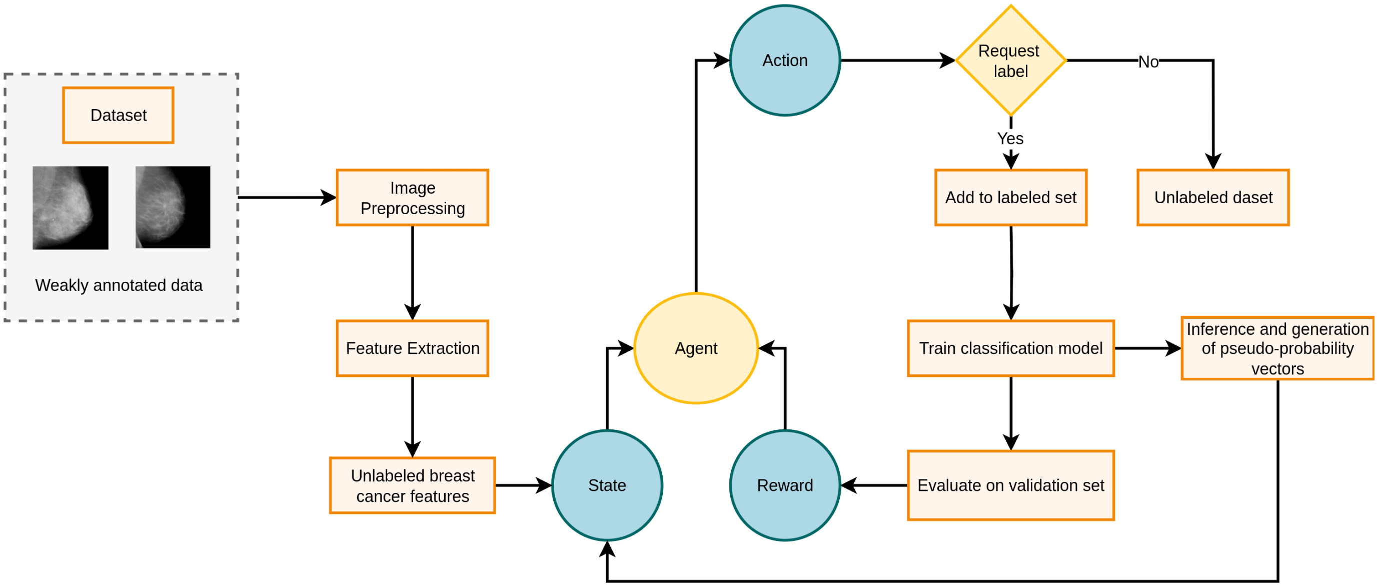
Dataset
Weakly annotated data
Image Preprocessing
Feature Extraction
Unlabeled breast cancer features
State
Agent
Action
Request label
No
Yes
Add to labeled set
Unlabeled daset
Train classification model
Inference and generation of pseudo-probability vectors
Evaluate on validation set
Reward

Figure 2. Process diagram of the Reinforcement Learning-based active sample selection for breast cancer image classification. The diagram illustrates the iterative cycle of sample selection, model training, and performance evaluation, highlighting the interaction between the RL agent, the unlabeled dataset, and the classification model.

IV. Results

A. Baseline Model

Table I summarizes the hyperparameters selected for the baseline model, which was developed using conventional supervised learning with hyperparameter tuning on the fully labeled dataset. This configuration achieved an F-score of 71%. Figure 3 presents the corresponding confusion matrix for the test set, in which approximately 80% of malignant cancer cases were correctly classified.

Table I
Final model hyperparameters and training configuration

Hyperparameter	Value / Description
Base Model	InceptionV3 pretrained on RadImageNet
Fine-Tuning Strategy	Trainable layers: mixed9 and mixed10; remaining layers frozen
Batch Size	256
Initial Learning Rate	10^{-3}
Learning Rate Scheduler	ReduceLROnPlateau, with factor = 0.5 (lr' = lr × factor) and patience = 5
Optimizer	Adam
Regularization	L2 kernel regularizer ($\lambda = 10^{-3}$) on dense layer
Dropout	0.4 after batch normalization
Noise Injection	Gaussian noise ($\sigma = 0.01$) after global average pooling
Data Augmentation	Horizontal flip; zoom = 0.1
Input Normalization	Rescale to $[-1, 1]$; $x \leftarrow \frac{x}{127.5} - 1.0$
Early Stopping	Monitor validation loss; patience = 50 epochs

We note that mammography alone does not provide a definitive diagnosis or a complete staging of breast cancer. While mammographic images offer insights, they capture only a fragment of the broader clinical picture, which typically involves additional clinical, laboratory, and histopathological information. Consequently, model performance should be interpreted with consideration for the limitations inherent in relying exclusively on imaging data for a complex diagnostic task.

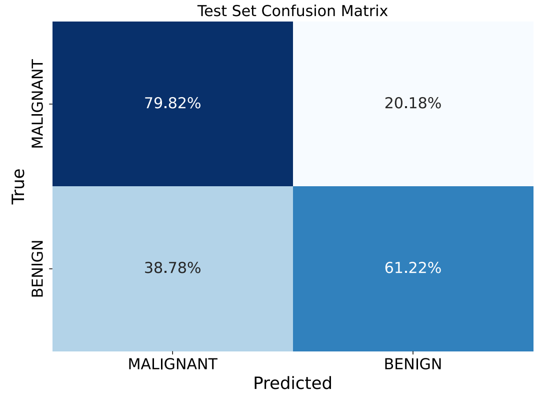

Figure 3. Confusion matrix of the baseline model on the test set.

B. Reinforcement Learning

At the end of the RL training process, the model achieved an F-score of 70%, closely approaching the baseline performance of 71%, while relying on a substantially smaller number of labeled samples. From a statistical standpoint, this demonstrates that the model was able to generalize effectively even with limited supervision. It is possible that, with continued training and access to a larger unlabeled pool, the model would surpass the baseline performance.

Figure 4 compares the model's performance in breast cancer prediction on the test set as new samples are progressively labeled.

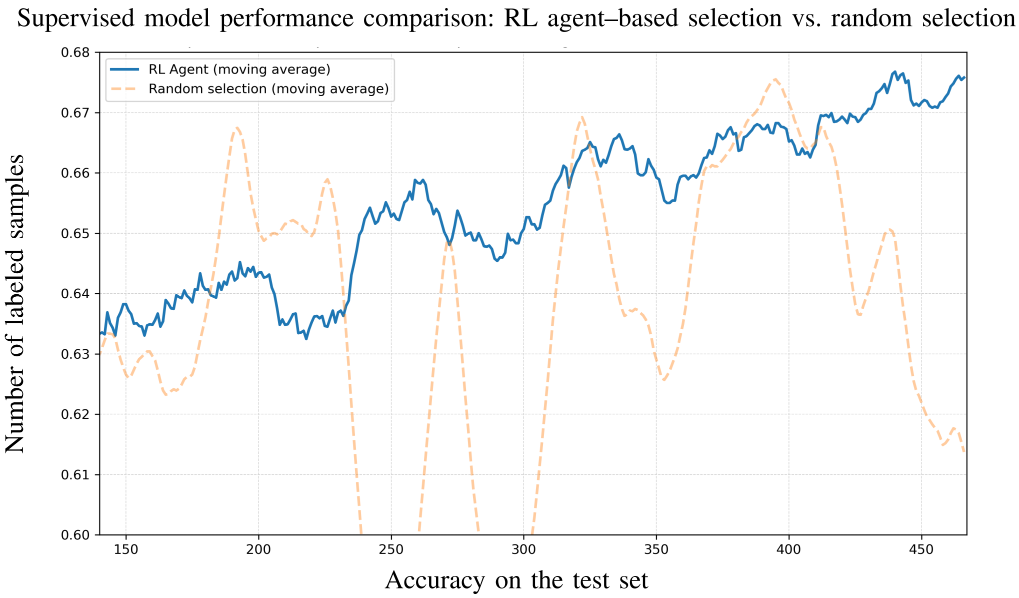

Figure 4. Performance comparison of breast cancer prediction models with progressive sample labeling. The blue curve shows the RL agent's selection strategy, while the orange dashed curve represents random sample selection. The RL–based approach demonstrates a consistent upward trend, indicating more effective identification of informative samples for annotation.

The blue curve, corresponding to the RL agent's selection strategy, demonstrates an upward trend in accuracy, indicating that the samples chosen by the agent tend to be more informative for the model. Furthermore, the model's performance did not exhibit a saturation trend, suggesting that the agent's policy could benefit from continued training and access to a larger pool of unlabeled samples.

In contrast, the orange dashed curve, representing random sample selection, shows a more unstable behavior, with fluctuations and no clear improvement pattern when only a few samples are added. Thus, without an informed policy, it is likely that less useful samples will be chosen, which does not contribute to learning.

To better understand the agent's selection behavior, Figure 5 illustrates the class distribution within the initially labeled set and the final set of samples selected by the RL agent. The initial annotations were obtained via stratified random sampling, simulating expert–provided labels to seed the learning process. Despite operating in a label–free setting, the RL–based selection policy maintained coverage of both classes in the final labeled subset. This indicates that the agent's sampling behavior remained consistent, suggesting a label–agnostic yet coherent acquisition strategy throughout the learning process.

These results highlight the effectiveness of the RL–based strategy in identifying more informative samples for annotation. This suggests potential for further performance gains in practical deployment scenarios, contingent upon the availability of additional resources, such as annotation time or budget.

Our approach shares similarities with other recent developments in the field. For instance, [19] demonstrated that uncertainty–based AL required only 26 labeled slices to achieve significant gains in lesion detection sensitivity. Similarly, our RL–based selection mechanism maintained representative class distributions and improved performance using a small fraction of the total label space. While our experiment started with a small stratified subset (5% of the samples), improvements could be made by incorporating more structured cold–start strategies, such as uncertainty sampling or clustering, which could guide the RL agent toward more effective initial policies and reduce the phase in which suboptimal actions may predominate. Despite these potential enhancements, our study successfully address the goal of minimizing expert annotation while preserving model performance.

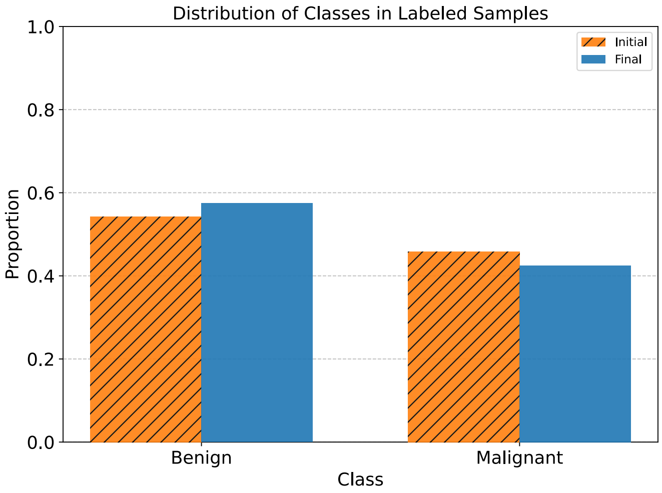

Figure 5. Class distribution in the initially labeled dataset (hatched bars) and the final set of samples labeled after agent–driven selection (non–hatched bars).

V. Conclusion

This work explores the use of RL for sample selection in labeling mammographic images, addressing the challenge of limited annotation resources in the medical field. The proposed RL–based strategy shows promise in identifying more informative samples for annotation, potentially reducing labeling costs in resource–constrained settings. Although it presents limitations in relying solely on mammographic images for cancer classification without additional clinical information, the study contributes by addressing the problem of efficient sample selection in supervised learning.

The approach offers opportunities for future developments, including integration with complementary clinical data, such as patient history, biomarkers, and pathology reports, to improve diagnostic accuracy. Further validation in real clinical scenarios is important to assess the applicability and effectiveness of the method. By prioritizing the selection of informative and high–quality samples, this research supports the development of more generalizable models that can be trained with fewer annotated examples. This is particularly relevant in multi–institutional settings, where data collection and expert labeling are expensive and heterogeneous.

References

[1] K. Yu and X. Xie, "Predicting hospital readmission: A joint ensemble-learning model," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 447–456, 2020.

[2] P. Fransvea, G. Fransvea, P. Liuzzi, G. Sganga, A. Mannini, and G. Costa, "Study and validation of an explainable machine learning-based mortality prediction following emergency surgery in the elderly: A prospective observational study," *International Journal of Surgery*, vol. 107, p. 106954, 2022.

[3] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning in medical image analysis," *Medical Image Analysis*, p. 103201, 2024.

[4] G. Lan, S. Xiao, J. Yang, J. Wen, W. Lu, and X. Gao, "Active learning inspired method in generative models," *Expert Systems with Applications*, vol. 249, p. 123582, 2024.

[5] M. Otten, A. R. Jagesar, T. A. Dam, L. A. Biesheuvel, F. den Hengst, K. A. Ziesemer, P. J. Thoral, H.-J. de Grooth, A. R. Girbes, V. François-Lavet *et al.*, "Does reinforcement learning improve outcomes for critically ill patients? a systematic review and level-of-readiness assessment," *Critical care medicine*, vol. 52, no. 2, pp. e79–e88, 2024.

[6] A. Yala, P. G. Mikhael, C. Lehman, G. Lin, F. Strand, Y.-L. Wan, K. Hughes, S. Satuluru, T. Kim, I. Banerjee *et al.*, "Optimizing risk-based breast cancer screening policies with reinforcement learning," *Nature medicine*, vol. 28, no. 1, pp. 136–143, 2022.

[7] R. Gonzalez Lopez, "Reinforcement learning in oncology: A comprehensive review," 2024.

[8] C.-Y. Yang, C. Shiranthika, C.-Y. Wang, K.-W. Chen, and S. Sumathipala, "Reinforcement learning strategies in cancer chemotherapy treatments: A review," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107280, 2023.

[9] J. C. Lauffenburger, E. Yom-Tov, P. A. Keller, M. E. McDonnell, K. L. Crum, G. Bhatkhande, E. S. Sears, K. Hanken, L. G. Bessette, C. P. Fontanet *et al.*, "The impact of using reinforcement learning to personalize communication on medication adherence: findings from the reinforce trial," *npj Digital Medicine*, vol. 7, no. 1, p. 39, 2024.

[10] L. Bohang, N. Li, J. Yang, O. Alfarraj, F. Albelhai, A. Tolba, Z. A. Shaikh, R. Alizadehsani, P. Pławiak, and P. L. Yee, "Image steganalysis using active learning and hyperparameter optimization," *Scientific Reports*, vol. 15, no. 1, p. 7340, 2025.

[11] C. Barata, V. Rotemberg, N. C. Codella, P. Tschandl, C. Rinner, B. N. Akay, Z. Apalla, G. Argenziano, A. Halpern, A. Lallas *et al.*, "A reinforcement learning model for ai-based decision support in skin cancer," *Nature Medicine*, vol. 29, no. 8, pp. 1941–1946, 2023.

[12] Y. Lai, J. Zhong, M. Li, S. Zhao, and X. Yang, "Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models," *arXiv preprint arXiv:2503.13939*, 2025.

[13] A. M. M. Kasmaee, A. Ataei, S. V. Moravvej, R. Alizadehsani, J. M. Gorriz, Y.-D. Zhang, R.-S. Tan, and U. R. Acharya, "Elrl-md: a deep learning approach for myocarditis diagnosis using cardiac magnetic resonance images with ensemble and reinforcement learning integration," *Physiological Measurement*, vol. 45, no. 5, p. 055011, 2024.

[14] S. Muksimova, S. Umirzakova, S. Kang, and Y. Im Cho, "Cervilearnnet: Advancing cervical cancer diagnosis with reinforcement learning-enhanced convolutional networks," *Heliyon*, vol. 10, no. 9, 2024.

[15] Y. Liu, D. Yuan, Z. Xu, Y. Zhan, H. Zhang, J. Lu, and T. Lukasiewicz, "Pixel level deep reinforcement learning for accurate and robust medical image segmentation," *Scientific Reports*, vol. 15, no. 1, p. 8213, 2025.

[16] S. Job, X. Tao, L. Li, H. Xie, T. Cai, J. Yong, and Q. Li, "Optimal treatment strategies for critical patients with deep reinforcement learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–22, 2024.

[17] P. Jayaraman, J. Desman, M. Sabounchi, G. N. Nadkarni, and A. Sakhuja, "A primer on reinforcement learning in medicine for clinicians," *NPJ Digital Medicine*, vol. 7, no. 1, p. 337, 2024.

[18] M. Santos and G. Marreiros, "A systematic review of active learning approaches in the selection of medical images," *Procedia Computer Science*, vol. 256, pp. 843–851, 2025.

[19] J. Huang, N. Farpour, B. J. Yang, M. Mupparapu, F. Lure, J. Li, H. Yan, and F. C. Setzer, "Uncertainty-based active learning by bayesian u-net for multi-label cone-beam ct segmentation," *Journal of Endodontics*, vol. 50, no. 2, pp. 220–228, 2024.

[20] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. Yuille, and Z. Zhou, "Making your first choice: to address cold start problem in medical active learning," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 496–525.

[21] J. Chen, B. Ma, H. Cui, and Y. Xia, "Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 439–11 449.

[22] H. Shahzad, A. Sanaullah, S. Arora, U. Drepper, and M. Herbordt, "Autoannotate: Reinforcement learning based code annotation for high level synthesis," in *2024 25th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2024, pp. 1–9.

[23] Y. Yang, W. Chen, Z. Li, Z. He, and M. Zhang, "Distantly supervised ner with partial annotation learning and reinforcement learning," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2159–2169.

[24] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[25] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of digital database for screening mammography (cbis-ddsm)," 2016.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[28] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 0, no. ja, p. e210315, 0. [Online]. Available: https://doi.org/10.1148/ryai.210315

[29] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," *arXiv preprint arXiv:1205.2653*, 2012.

[30] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulao, A. Kallinteris, M. Krimmel, A. KG *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2407.17032*, 2024.

[31] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.