OPTIMAL BIOPSY DECISION MAKING IN BREAST CANCER USING REINFORCEMENT LEARNING

By

Ruchi Parmar, B. Tech., Nirma University, 2018

A Major Research Paper
presented to Toronto Metropolitan University
in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2024

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions, as accepted by my examiners.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Ruchi Parmar

OPTIMAL BIOPSY DECISION MAKING IN BREAST CANCER USING REINFORCEMENT LEARNING

Ruchi Parmar

Master of Science 2024

Data Science and Analytics

Toronto Metropolitan University

ABSTRACT

In this Major Research Project, I applied various Reinforcement Learning(RL) methods to optimize biopsy decision-making for breast cancer. The main objective of the research is to minimize the number of False Positive biopsies while improving patient outcomes. I started with implementing a traditional Markov Decision Process(MDP) based approach using a Backward Induction algorithm to generate policies for different age-states. Later I implemented advanced RL approaches, Q - Learning, and Deep Q-Network(DQN) to create policies and compare results. Generated policies are further tested and compared using simulation and publicly available datasets. The results shows that while Backward Induction and Q-Learning performed well and gave the desired results, DQN model struggles to learn the pattern. This methods can further be implemented and tested along with large clinical data and cancer probabilities generated using CAD models used in practice.

Keywords:

Reinforcement Learning, Markov Decision Process, Backward Induction, Q-Learning, Deep Q-Learning, Breast Cancer, Biopsy, Policy

ACKNOWLEDGEMENTS

I am extremely grateful to Professor **Dr. Mucahit Cevik** for the guidance and supervision I received throughout my Major research project(MRP). His expertise in the field of Reinforcement Learning and Breast Cancer optimization research helped me a lot to learn and implement various solutions for biopsy decision-making optimization. His patience and insightful feedback throughout the research and implementation played a crucial role in the successful completion of the project. I am grateful for his encouragement and support, which have significantly contributed to my academic and personal growth during this project.

Thank you Professor Dr. Cevik for your valuable time, guidance and for being an inspiring mentor.

Table of Contents

List of Figures

List of Tables

1. Introduction

This document provides the details on the domain, topic and research questions of the MRP. It begins with a brief background on the topic and datasets, defines the problem, and states the research question in Introduction section. This is followed by a literature review and a detailed exploratory analysis of the dataset. Next I have outlined the methodology and experiments performed to develop Reinforcement Learning algorithms, go over the results and finish the paper with a discussion about this project and recommendations for future work.

A. Background

Breast Cancer is a major health challenge across the world, with increasing number of cases. It is mentioned that 2.3 million new cases of breast cancer were diagnosed in 2022 and there were 670,00 total deaths globally[34]. Annual Mammography is the standard practice followed by most of the advanced countries to detect Breast Cancer in early stages. In the United States approximately 39 million mammograms are performed annually as of year 2021[12]. Post mammogram more than 1 million women undergo breast biopsies annually. These biopsies are crucial for diagnosing breast abnormalities and determining whether a lump or mass is benign or cancerous. Approximately 20% of breast biopsies performed in the U.S. each year turn out to be cancerous, while the remaining 80% are false positives (benign)[22]. This cause unnecessary treatments, patient anxiety, and expenditures. This project focuses on exploring Reinforcement Learning(RL) techniques to optimize breast cancer biopsy decision-making. This approach focuses on optimizing decision-making, reducing unnecessary procedures, and to mitigate the economic impact of false positive cases, ultimately improving patient outcomes. I have utilized data of State-Transition probabilities for breast cancer risks available on GitHub Repository[9] to build RL framework.

B. Research Objectives

Through this major research project, I tried to achieve several objectives in two major parts. First, I implemented traditional Markov Decision Process(MDP) based Reinforced Learning method for biopsy decision-making and obtain results with available data in line with existing research. Next, I applied advanced reinforcement learning techniques, Q-Learning and Deep Q-Networks, to further optimize the biopsy decision-making process. I compared the impact of these

methods on reducing number of false positive cases(unnecessary procedures) while maintaining number of true positive cases(actual cancer cases) and to compare the effectiveness of these techniques against traditional MDP-based(Markov Decision Process) methods. Additionally, I formulated reward functions to guide RL agents in making decisions, considering factors such as the disutility of biopsy, to achieve better performance.

2. Literature Review

The primary focus of this project is to understand the approaches used by radiologists in making decisions based on mammogram screenings and to replicate traditional reinforcement learning (RL) methods for training an RL agent to optimally make biopsy decisions while balancing false positives (FP) and true positives (TP). Next, utilize advance RL methods to further optimize biopsy decision making process. Although several studies have explored traditional RL methods in this domain, applying Q-Learning and Deep Q-Learning for optimizing biopsy decision-making is a novel approach.

In this section, I provide an overview of the articles referenced for this project. I began by reviewing articles related to traditional methods for detecting breast cancer. Regular mammography screening is the standard practice for detection of cancer at early stage in asymptomatic women. Organizations such as the American Cancer Society (ACS) advocate mammography as the primary method for identifying breast cancer. The ACS advises that women undergo clinical breast examinations every three years between the ages of 20 and 40 and annually after age 40. For women at average risk, screening mammography is recommended annually starting at age 40 and should continue as long as the woman is eligible for basic breast cancer treatment (Smith et al. (2010) [27]). This practice has been shown to significantly reduce breast cancer mortality (Nelson et al. (2009) [23]).

After screening mammography, radiologists evaluate the likelihood of cancer based on the findings and decide on the appropriate action, balancing the need for early cancer detection using biopsy against the risk of false positives. Depending on the patient's cancer risk, radiologists typically choose one of three options: (1)routine follow-up mammography in a year, (2)short-term follow-up mammography in six months, or (3)immediate diagnostic actions such as biopsy (Ayvaci and Burnside (2012) [3]).

Mammogram interpretation can go wrong in two ways. First, a mammogram can be interpreted

as normal when cancer is present, leading to delayed diagnosis and increased risk to the patient's life. While some false negatives are due to cancers that are not visible on the mammogram, the majority are cancers that could have been detected in hindsight (Baines and Dayan (1999) [4], Bird, Wallace, and Yankaskas (1992) [8]). Second, a mammogram can be incorrectly labeled as positive when there is no cancer, resulting in over-treatment and unnecessary anxiety for patients. The false-positive rate of mammography is nearly 80% in United States ([22]). These error rates highlight the complexity of interpreting mammograms.

The accuracy of mammogram reading differs with the radiologist's skills and training (Barlow et al. (2004) [6], Beam, Layde, and Sullivan (1996) [7]). Radiologists, with fellowship training in mammography and extensive focus on this domain, generally perform better than general radiologists, who interpret mammograms as part of a broader practice (Sickles, Wolverton, and Dee (2002) [26]). Additionally, there are performance differences between countries also. Research found that while cancer detection rates are similar in the U.S. and the U.K., U.S. radiologists labeled many more mammograms as uncertain or suspicious, leading to at least twice as many follow-up tests, such as biopsies (Smith-Bindman et al. (2003) [28]).

To ensure consistency in mammography practices, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon (BI-RADS (1998) [25], Liberman and Menell (2002) [20]). This system includes codes as key indicators for benign or malignant diagnoses. It assists radiologists and referring physicians in making informed breast cancer decisions and managing patients. Each mammogram is described using standardized descriptors and categorized into one of six final assessment categories (BI-RADS codes) based on the mammogram's observations as seen in Table 1. For instance, BI-RADS 0 requires additional imaging; BI-RADS 1 and 2 suggest no further action and continued routine screening for no or benign findings; BI-RADS 3 indicates a probably benign finding needing short-term follow-up (typically six months to two years); and BI-RADS 4 and 5 suggest suspicious or high-risk findings, recommending an immediate biopsy (Chhatwal, Alagoz, and Burnside (2010) [14]).

Despite this standardization, the current use of the BI-RADS lexicon has many issues. Notably, the BI-RADS categories do not account for age in biopsy decision-making, despite older age groups having specific characteristics that need particular consideration. For example, breast cancers in older women are often less aggressive, implying that a higher action threshold might be suitable (Fowble et al. (1994) [17]). Also, false-positive mammograms that lead to unnecessary invasive procedures can be more challenging or risky for older individuals with

Table 1: BI-RADS assessment codes

BI-RADS	Explanation	Recommended Action
0	Need additional Imaging evaluation	additional Imaging evaluation
1	Negative finding	routine yearly screening
2	Benign Finding	routine yearly screening
3	Probably Benign (less than 2% risk of cancer)	Short-term follow-up (6 months)
4	Suspicious (risk of cancer is between 2% and 95%)	Biopsy
5	Highly suggestive of malignancy (95% risk of cancer)	Biopsy

other health conditions. These unique features of breast cancer in older women suggest that biopsy probability thresholds might differ from those in younger women (Chhatwal, Alagoz, and Burnside (2010) [14]).

Computer-aided diagnostic (CADx) models show promise to enhance early breast cancer diagnosis as per different researches. For example, Few models like Bayesian network to predict malignancy risk using mammographic features and demographic data (developed by Burnside et al. (2006) [11], Burnside et al. (2009) [10]), artificial neural networks (used by [5] and Ayer et al. (2010) [2]), decision fusion(used by Jesneck et al. (2006) [19]) and logistic regression (used by Chhatwal et al. (2009) [15]) performed better and gave more accurate results. This models generally provides probability of cancer instead of BI-RADS categories, which can be more helpful to make accurate decisions. However, this methods also doesn't provide any threshold for biopsy decision with respect to patient related factors such as age, cost etc.(Chhatwal, Alagoz, and Burnside (2010) [14]).

These methods doesn't help to determine an optimal threshold for recommending biopsies to reduce high number of false-positives. Furthermore, false-positive mammograms cause unnecessary anxiety, pain, and potential complications for patients. Biopsies can also introduce changes (such as distortion) in future mammograms and complicates future diagnoses ([14]). There are several studies which utilised traditional Markov Decision Based Reinforcement Learning methods for various use cases such as finding optimal biopsy decision making, to find optimal screening intervals, to find screening polices under various constraints such as cost or to reduce the over diagnosis.

Reinforcement learning (RL) (Sutton and Barto (2018) [31], Szepesvári (2010) [32], Powell

(2011) [24], Hengst et al. (2020) [18], Dong, Ding, and Zhang (2020) [16]), is a type of adaptive mechanism and a sub-field of machine learning where agents learn dynamically through heuristic methods by interacting with their environment. Unlike other computational approaches, RL allows agents to learn without needing specific supervision or a complete model of the environment. Typically, an agent's actions in the environment result in both immediate and delayed rewards. Agent searches for the optimal next action within the state space. The agent operates through a trial-and-error process to maximize the total accumulated rewards over time (Yang et al. (2023) [35]).

In one of the studies, author developed Optimal Biopsy Decision Model(OBDM) that integrates both mammographic and Demographic features using finite-horizon discrete time Markov Decision Process(MDP). With nearly 65,892 clinical data, model parameters such as state transition probabilities were estimated. This model recommends fewer biopsies compared to radiologists, reducing false positives without significantly increasing false negatives. Age-dependent biopsy thresholds indicate higher thresholds for older women, aligning with less aggressive cancer progression and comorbidities in this group (Chhatwal, Alagoz, and Burnside (2010) [14]).

In another paper, author utilised large-scale finite-horizon Markov Decision Process (MDP) model incorporating 4.6 million states to reduce number of over diagnosis in breast cancers which leads to unnecessary treatments and healthcare costs. This model aims to distinguish between invasive and indolent breast cancers. It employed divide-and-search algorithm for optimal decision threshold identification. This resulted into potential 20% reduction in over diagnosis through tailored diagnostic guidelines. This study also found that age-specific diagnostic policies could result in substantial quality assured life years(QALY) gains and cost savings and identified that current practices could lead to aggressive diagnostic actions, especially in older women (Tunç, Alagoz, and Burnside (2022) [33]).

Other study aimed to assess the value and effectiveness of dynamic mammography screening policies over static ones. Author Created a partially observed Markov chain model to emulate breast cancer progression and investigate dynamic policies that adjust screening intervals based on age-related risk factors. By analyzing wide range of two-phase screening policies for their impact on lifetime mortality risk, author concluded that dynamic, two-phase screening policies could be more beneficial than the current recommendations (Maillart et al. (2008) [21]). A. Ayer and Stout used a Partially Observable Markov Decision Process (POMDP) to optimize screening schedules. They developed a personalized screening model that considers a broader range of individual risk factors such as personal screening history. They tried to improve quality-adjusted

life years(QALYs) and reduce unnecessary screenings. This study concluded that personalised screening strategies could potentially lead to fewer mammograms and false positives while increasing QALYs, especially in high-risk groups (Ayer and Stout (2012) [1]).

In another study, author focused on optimizing diagnostic decisions after mammography, maximizing Quality-Adjusted Life Years (QALYs) within budget constraints. They created a finite-horizon constrained Markov Decision Process (MDP) model. Results showed that Optimal decision-making could lead to approximately 22% savings in costs without compromising QALYs. Short-term follow-ups identified as primary targets for cost reduction when there is budget constraints. The study's findings highlight the importance of integrating cost considerations with health outcomes to improve healthcare decision-making (Ayvaci and Burnside (2012) [3]). Cevik et al. focuses on optimal allocation of limited mammography resources for effective breast cancer screening. Early detection of breast cancer through mammography is crucial but, limited resources (including cost and availability of trained personnel and diagnostic machines) hinder widespread adoption of mammography screening. The study proposes a constrained partially observable Markov decision process (CPOMDP) model. It concludes that efficient allocation of resources based on different risk levels significantly improves QALY gains, especially for patients at higher breast cancer risk. As mammography screening capacity decreases, patients in the 40–49 age group receive the least priority for screening (Cevik et al. (2018) [13]).

Most of these studies used traditional MDP based approaches. To the best of my knowledge, no study has explored the use of advanced reinforcement learning methods, such as Q-Learning or Deep Q-Networks(DQN) specifically for the Optimal Biopsy Decision Making. As many other studies in healthcare domain have used these methods, they have observed better performances. Traditional RL methods follows model/algorithm based approach, whereas Q-learning and DQN follows learning from environment by following action-reward strategy. One of the study on brain tumor localization study addresses limitations in AI for radiology, including large annotated datasets, non-generalizability, and inadequate explainability and explores reinforcement learning (RL) as a solution for these challenges. They trained Deep Q Network and incorporated image exploration with rewards and punishments to localize lesions. This study observed that RL predictions consistently improved during training and achieved 85% accuracy in predicting lesion locations outperforming supervised deep learning(7% accuracy) (Stember and Shalu (2020) [30]).

This research project focus on using these techniques to find threshold for biopsy recommendation in Breast cancer for Optimal Biopsy Decision making and compare results with transitional methods.

3. Exploratory Data Analysis

This section aims to provide a comprehensive understanding of the dataset used for the research project. It contains the information on Data Source and Files, description of data's basic features. It includes more detailed analysis, such as relation between variables, identifying trends. EDA helped to gain a deeper understanding of the data.

A. Data Source and Files

Data for this project is acquired from Github open Repository - Breast Cancer transition Probabilities[9]. The datasets are open-sourced and compliant with the MRP requirements. It contains two csv files -

- "all_women_data.csv" - This file contains data of 10000 women's probability of cancer at early stage and advanced stage at each time stamp(age 40-100). Patients start screening at age 40 and goes under screening each year as recommendation.
- "state_transition_probabilities.csv" - This file contains state transition probabilities at each age for RL model for Regular Mammogram and Biopsy actions. It considers cancer risk score as a state of RL model(i.e. MDP) and calculate transition probabilities of states at different age.

Data files are available on GitHub. Link to the GitHub repository is included in References.

B. Data Analysis

In this section we will analyze we will analyze each feature of both data files, exploring their individual properties and interrelationships. Additionally, we will examine the rewards distribution for each action (Annual Mammogram & Biopsy) of the RL agent used in the project.

- Patient ID - ID of patient(10000 women in total). This column has no nulls.
- Age - Age of patient at time of screening. Age ranges between 40 to 100. As per the guidelines women usually starts regular screening at the age of 40.

- Stages - There are total 5 features according to 5 different stages of cancer and death - Stage 0 - Healthy, Stage 1 - Early stage cancer, Stage 2 - Advanced stage cancer, Stage 3 - Death from cancer, Stage 4 - Death from other causes. These features contains values of the probability of patient being in each stage. i.e. if probability if patient being healthy at given age.

  These probabilities ranges between 0 to 1. In the following graphs(Figure 1 & Figure 2) we can see the distribution of probabilities of patient being in early or advanced stage cancer.

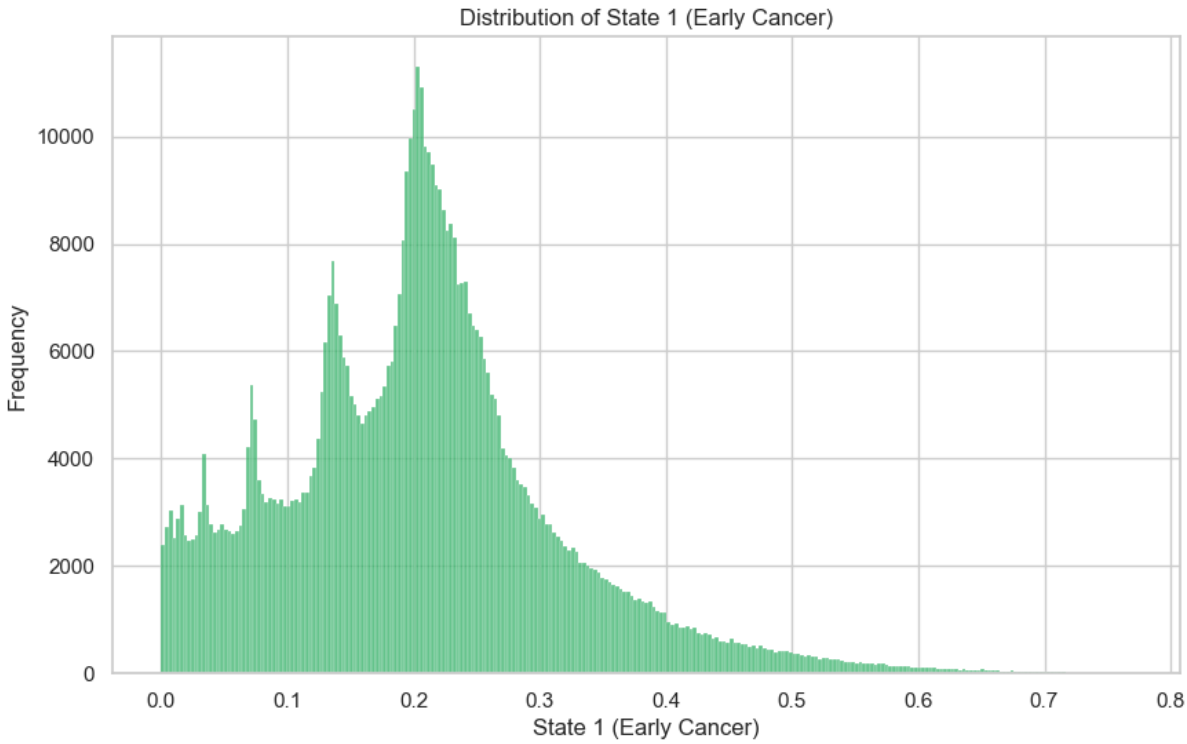


Figure 1: Distribution of Feature - State 0(Early Cancer)

- Risk Score - Risk score is the current probability of cancer for patient. It is calculated using [(1 - probability of being healthy) * 100]. The risk score is calculated by converting the probability of breast cancer in a number ranging from 0 to 100. For example, if the probability of patient x in state 0 - Healthy at age t is 0.7500, then probability of breast cancer is 0.2500, then the risk score for patient x at age t is 25.

  This Risk scores will be used in the RL algorithms as state space S, where S = 0, 1, 2,...., 100. Where S = 0 meaning patient has 0 probability of cancer. At any time t patient will be in state s, $s \in \mathbf{S}$.

- state transition probabilities - Data features From State, To state and Probability represents the state transition probabilities at each age from 40 to 100.

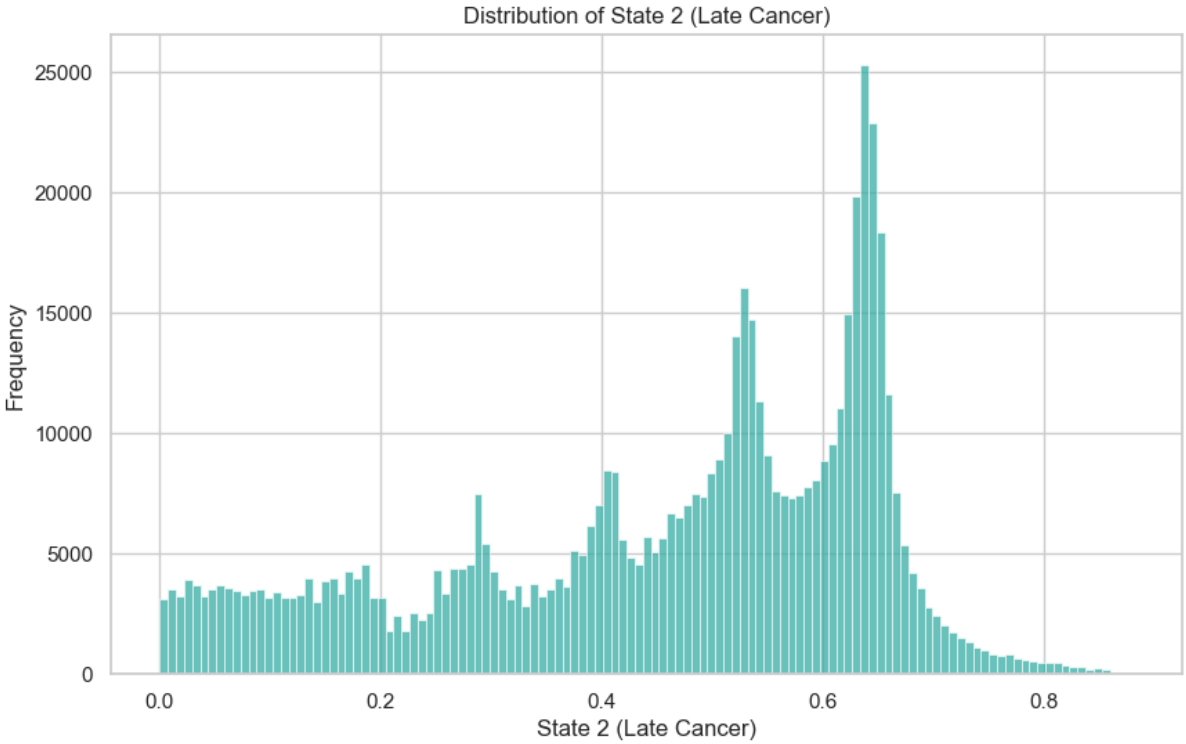


Figure 2: Distribution of Feature - State 1(Advanced Cancer)

For example, if patient at timestamp(age) t = 40 is in stage s1 = 5 (5% probability of breast cancer), then at next timestamp t + 1 = 41, patient being in any state s2, $s2 \in \mathbf{S}$ is probability P. In Figure 3, example of transition probabilities are given. This probabilities are taken from study by Maillart et al. (2008) [21]. This model has only 5 stages and they have probabilities for age groups such as 40 - 44, 45-49, etc.

$$\alpha \in [45, 49] \qquad \alpha \in [50, 54]$$

$$P_\alpha = \begin{pmatrix} 0.99784 & 0.00097505 & 0 & 0 & 0.0011804 \\ 0 & 0.79049 & 0.20833 & 0 & 0.0011804 \\ 0 & 0 & 0.87870 & 0.12011 & 0.0011804 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad P_\alpha = \begin{pmatrix} 0.99703 & 0.0012814 & 0 & 0 & 0.0016911 \\ 0 & 0.86317 & 0.13514 & 0 & 0.0016911 \\ 0 & 0 & 0.85732 & 0.14099 & 0.0016911 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 3: Example of sample transition probabilities

Same way, our data has transition probabilities at each age 40, 41,..., 100 from stat s1 to s2 where $s1, s2 \in \mathbf{S}$. These probabilities are available in "state_transition_probabilities.csv" file due to large scale.

- Rewards - In Reinforcement Learning, an RL agent receives rewards for its actions, aiming to maximize the cumulative rewards to gain the greatest overall benefit.

For this project, I have developed an RL agent with two possible actions: Annual Mammogram and Biopsy. At any age between 40 and 100, the agent receives rewards based on the probability of cancer (either early or advanced stage) and the action taken. These rewards are derived with reference from multiple studies (mentioned in the Literature Review) and general research on breast cancer and biopsies. The rewards are expressed in terms of the total expected years of life following each action. If the action is an Annual Mammogram, the patient receives a reward ranging from 1 year to a few months, depending on their age and cancer probabilities, until the next screening. If the action is a Biopsy, the patient receives a one-time lump sum reward in terms of the total expected life years after the biopsy.

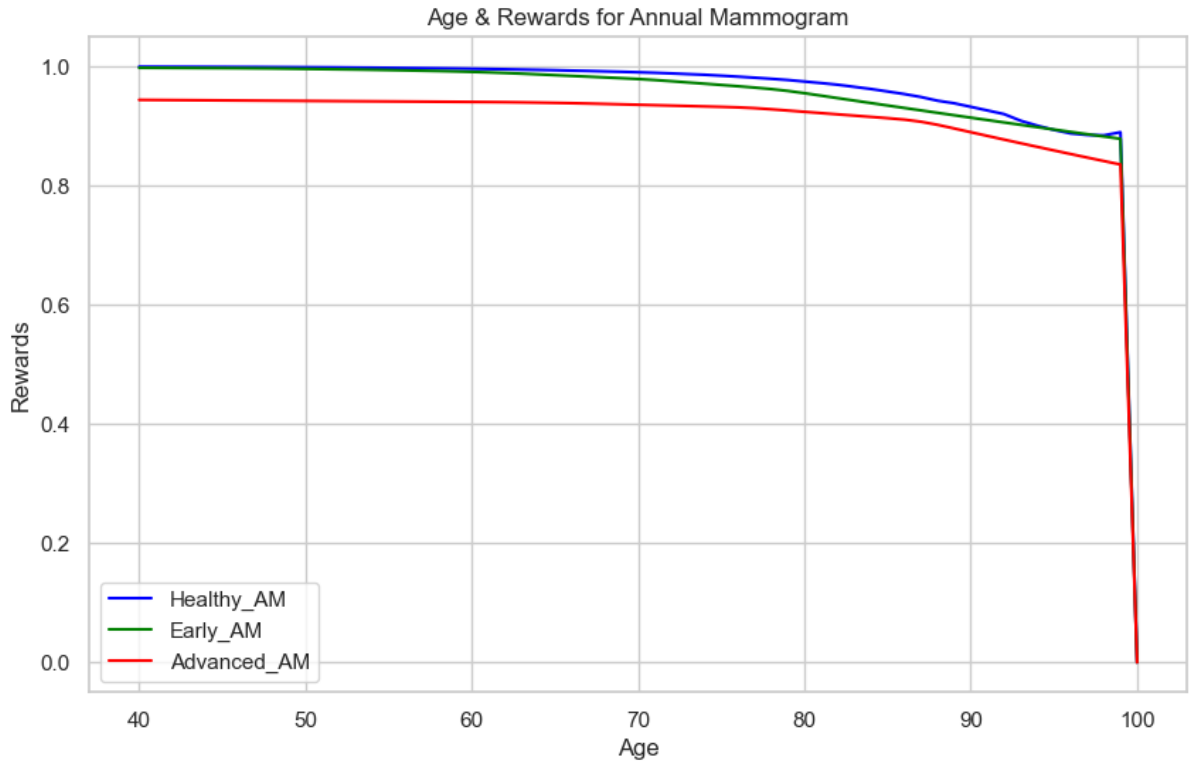


Figure 4: Age - Rewards for Healthy (s = 0), Early & Advance stages with action - Annual Mammogram

As shown in Figure 4 and 5, for the action Annual Mammogram, at a very young age and with a low probability of cancer, the reward is given as 1 year and decreases with age. Similarly, the rewards for Biopsy also decrease with age. The rewards also diminish as cancer probabilities (early stage or advanced stage) increase. Also, Early and Advance stages of cancer varies with different age groups. i.e., for very young women, small probability of cancer like 2% is advanced whereas with increase in age it can be higher[29].

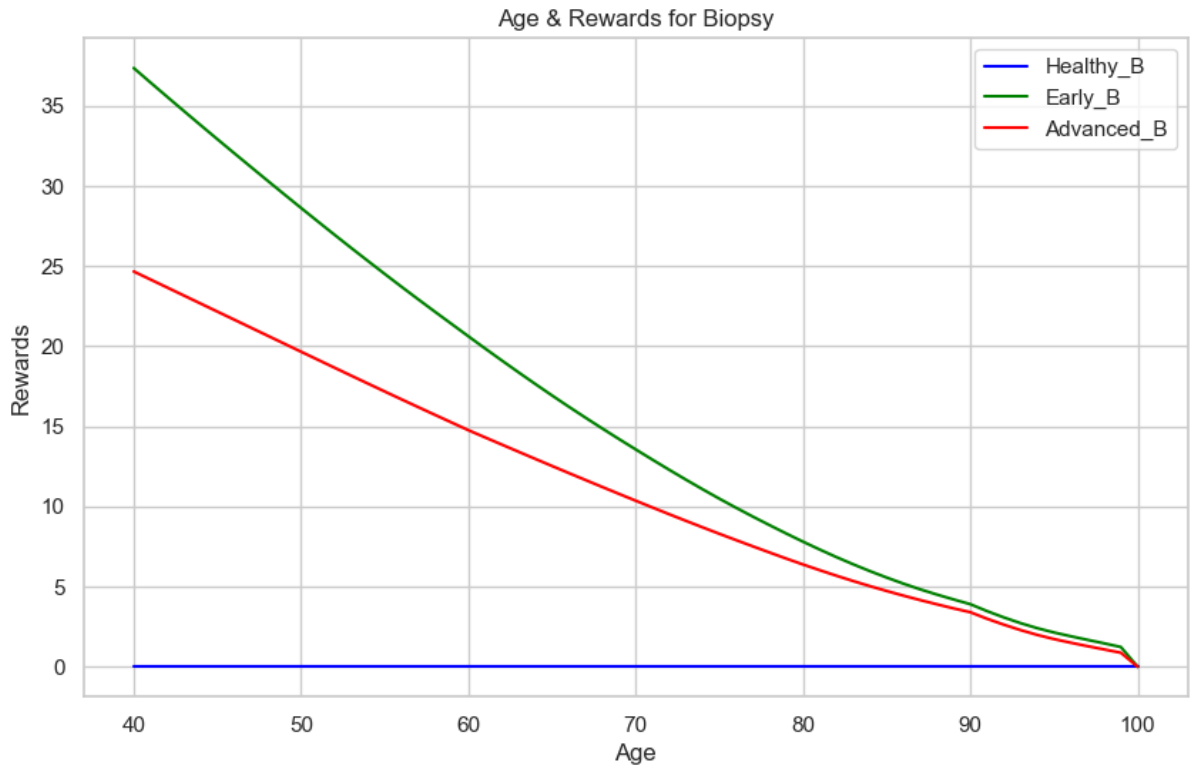


Figure 5: Age - Rewards for Healthy (s = 0), Early & Advance stages with action - Biopsy

C. Correlation between age and Probability of being Healthy

From the data, it is observed that probability of being Healthy decreases with increase in age. Same way, probability of cancer increases with increase in age. In Figure 6, it can be seen that If we take average probability of being healthy and cancer for each age, with age Chances of Being Healthy decreases and Chances of cancer increases. This figure has mean probability for patients from dataset at each age.

4. Methodology and Experiments

A. Aim of Study

The aim of this study is to develop and evaluate reinforcement learning (RL) algorithms to optimize breast cancer screening and diagnostic decision-making. Mainly the study focuses on determining the optimal actions(Annual Mammogram or Biopsy) at different risk levels and ages to maximize long-term health outcomes and minimize unnecessary procedures. different RL approaches like traditional MDP based Backward Induction and advanced method, Q-learning

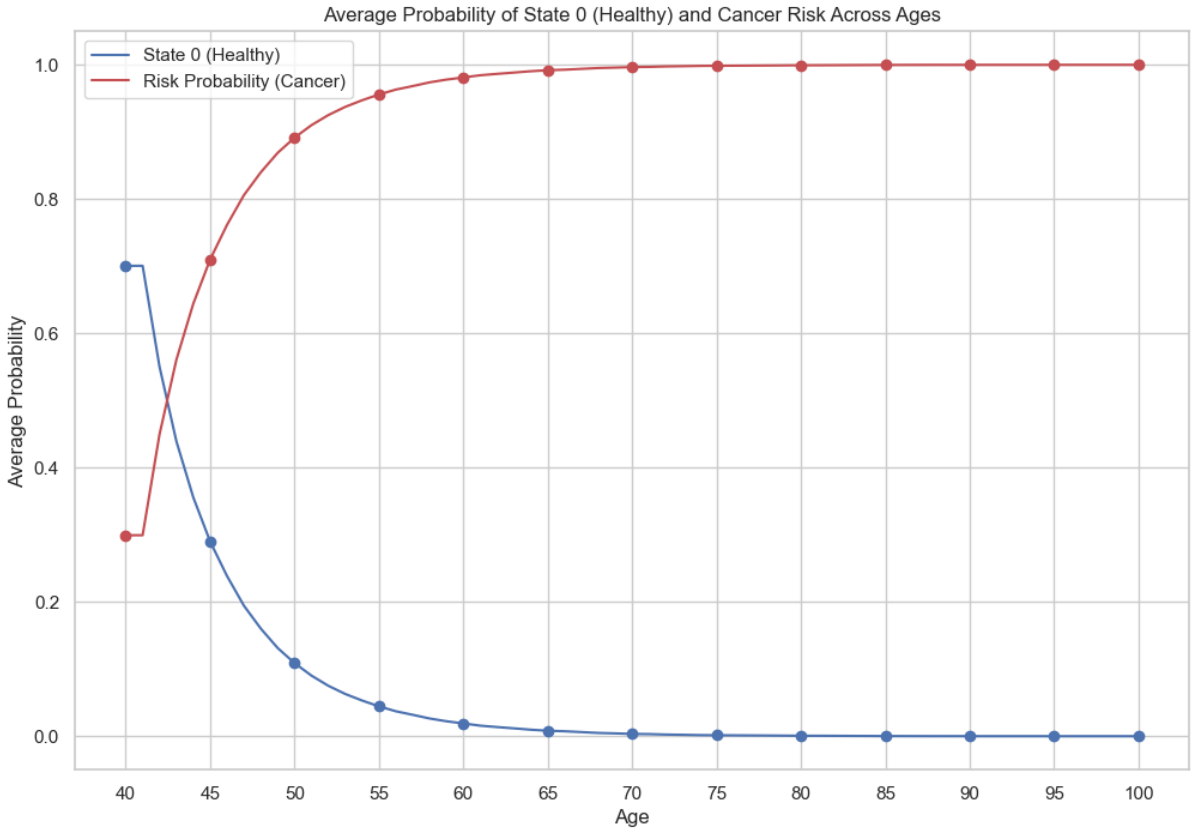


Figure 6: Average Probability of State 0 (Healthy) and Cancer Risk Across Ages

and DQN is implemented for the study.

B. Response(Dependent) and Independent Variable(s)

- **Response(Dependent) Variable**:
  - Long-term health outcome, measured by the reward function which incorporates the health states(Healthy, Early, Advanced, and Terminal) and the utility of diagnostic actions
- **Independent Variables**:
  - Age of the women(ranging from 40 to 99 years)
  - Risk score of the women(a metric representing the probability of having breast cancer)

C. Factors and Levels

In this experiment, the factors are the different RL methods being implemented and tested. Levels for the experiments are the optimization of Biopsy decision making while reduce unnecessary procedures, and mitigate the economic impact of false positives.

D. Experimental Design

The study includes three RL methods to develop and compare policies for optimal decision-making. These methods are Backward Induction, Q-Learning and Deep Q-Networks. The problem is framed using concepts of - Markov Decision Process(MDP) as seen in Figure 7 -

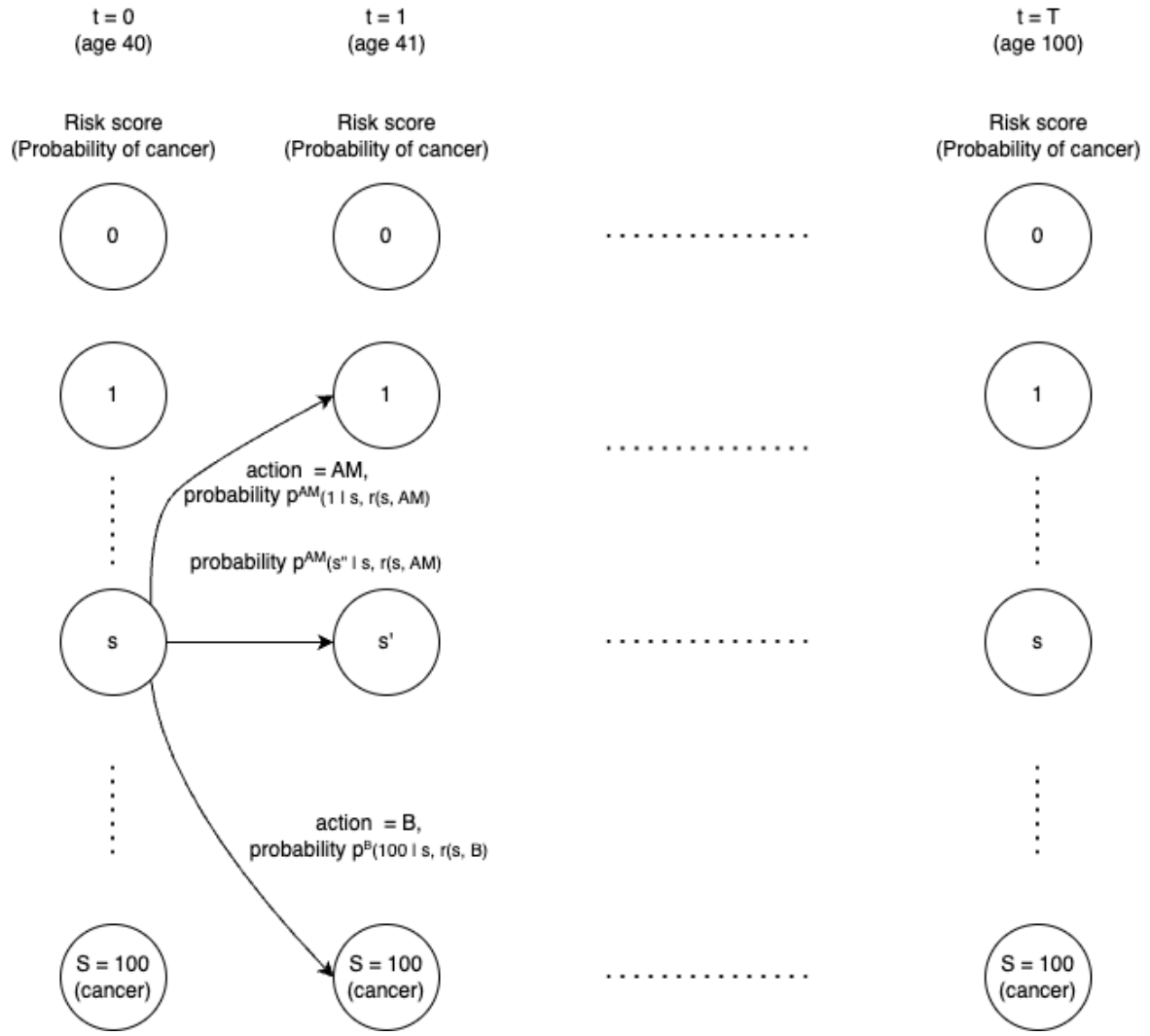


Figure 7: State transition diagram

- **States(S)**: Represent the health status of a woman, captured by the risk score and

categorized into different health states(Healthy, Early-stage Cancer, Advanced-stage Cancer, Terminal)

- **Actions(A)**: Possible actions, namely Annual Mammogram(AM) and Biopsy(B)
- **Transition Probabilities(P)**: Probabilities of moving from one state to another given a specific action
- **Rewards(R)**: Immediate rewards received after transitioning from one state to another, reflecting the health outcome and cost of actions

RL approaches implemented -

1. **Backward Induction**: Backward induction is a dynamic programming method used to solve MDPs by iteratively computing the optimal policy starting from the terminal state and working backward to the initial state.

   Algorithm Steps:

   (a) **Initialization**:
   - Define the set of states(S), actions(A), and time steps(T)
   - Initialize the utilities(U) for each state at the final time step(age 100) based on terminal rewards

   (b) **Backward Iteration**:
   - For each time step t from 99 to 40:
     - For each state s in S:
       * Compute the expected utility for each action a in A by considering the rewards and the expected future utilities
       * Select the action that maximizes the expected utility -

$$V_t(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_{t+1}(s') \right]$$

       * Update the utility of state s and the policy for time step t

   (c) **Policy Extraction**:
   - The resulting policy indicates the optimal action to take at each state and time step

   States:

- Risk scores ranging from 0 to 100
- Terminal state 100, which is an absorbing state indicating cancer detection

Rewards:

- Defined for each combination of state, action, and time step
- Example: Reward for 'AM' and 'B' actions in Healthy/ Early-stage cancer/ Advanced stage cancer state at given time stamp(age) t

2. **Q-Learning**: Q-learning is an off-policy temporal difference learning method used to find the optimal action-selection policy for an MDP by learning the action-value function Q(s,a) through interactions with the environment.

   Algorithm Steps:

   (a) **Initialization**:
   - Initialize Q-values for all state-action pairs
   - Set the learning rate(alpha) and discount factor(gamma)

   (b) **Epsilon-Greedy Policy**:
   - Implement an epsilon-greedy policy to balance exploration and exploitation. With probability epsilon, select a random action; otherwise, select the action with the highest Q-value.

   (c) **Action Execution**:
   - For each episode:
     - Start from the initial state
     - At each time step, select an action based on the epsilon-greedy policy
     - Execute the action, observe the reward and the next state
     - Update the Q-value for the state-action pair using the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

     - Transition to the next state and repeat until the terminal state is reached or the episode ends

   (d) **Policy Extraction**:
   - The optimal policy is derived by selecting the action with the highest Q-value for each state.

States: Same as backward induction, ranging from risk scores 0 to 100

Rewards: Same as backward induction, defined for each state-action pair

Exploration vs. Exploitation: The epsilon-greedy strategy ensures that the model explores new actions with probability epsilon and exploits the best-known actions with probability(1 - epsilon).

3. **DQN**: Deep Q-Network(DQN) enhances Q-learning by using a neural network to approximate the Q-values for state-action pairs, allowing it to handle large, continuous state spaces. It improves stability through techniques like experience replay and the use of a separate target network.

   Algorithm Steps:

   (a) **Initialization**:

   - **Neural Network Architecture**: Define a neural network with an input layer corresponding to the state space (risk scores ranging from 0 to 100) and output layer representing Q-values for each possible action (Annual Mammogram or Biopsy) and initialize the network weights.
   - **Experience Replay**: Implement an experience replay buffer to store transitions (s,a,r,s') and sample mini-batches for training the network, thus breaking correlations between consecutive samples.
   - **Target Network**: Use a target network to stabilize training by periodically updating its weights with the primary network's weights.

   (b) **Epsilon-Greedy Policy**:

   - Similar to Q-learning, employ an epsilon-greedy policy to balance exploration and exploitation, where the agent selects a random action with probability e and the best-known action otherwise.

   (c) **Action Execution**:

   - For each episode:
     - Start from the initial state.
     - Select an action based on the epsilon-greedy policy.
     - Execute the action, observe the reward, and transition to the next state.
     - Store the transition in the replay buffer.

- Sample a mini-batch from the replay buffer and perform gradient descent on the loss between the predicted Q-values and the target Q-values (calculated using the Bellman equation and the target network).
- Periodically update the target network's weights.

(d) **Policy Extraction**:

- After training, extract the optimal policy by selecting the action with the highest Q-value from the primary network for each state.

States: Same as Q-Learning, ranging from risk scores 0 to 100

Rewards: Same as Q-Learning, defined for each state-action pair

E. Experiment Performance and Revisions

The experiments are conducted in multiple iterations to refine the RL algorithms and improve the policies:

1. **Initial Setup**: Implement Backward Induction, Q-Learning, and Deep Q-Network (DQN) algorithms using the datasets. The initial methods are run with basic hyper-parameters and without any optimizations.
2. **Revisions**: Based on the initial results, several revisions were made:
   - Adjust the learning rate(alpha) and discount factor(gamma) for Q-learning
   - Refine the reward function to better capture the clinical significance of different health states
   - Add exploratory strategies(epsilon-greedy) to balance exploration and exploitation
   - Optimize the neural network architecture (e.g., number of layers, units).
   - Experiment with different activation functions, learning rates, and epsilon decay strategies to enhance model stability and convergence.

F. Measuring Performance

Policy Evaluation: The performance of the RL methods are measured by evaluating the learned policies.

G. Algorithm Comparison and Selection

The following criteria will be used to compare the backward induction, Q-learning and DQN algorithms:

- **Convergence**: Speed and stability of convergence to an optimal policy
- **Policy Quality**: Effectiveness of the policy in maximizing long-term rewards and health outcomes
- **Computational Efficiency**: Time and resources required for algorithms to get the policy

5. Results & Discussion

This section presents the results of the Optimal Biopsy Decision-Making for Breast Cancer using Reinforcement Learning methods, specifically Backward Induction and Q-Learning. It also discusses the insights observed from DQN implementation. The implementation was carried out in the following phases:

1. **Initial setup** - Implemented the Backward Induction algorithm as described in the literature review to generate a policy for different cancer stages at each time stamp (age), ranging from 40 to 100
2. **Q - Learning** - Applied the advanced reinforcement learning method, Q-Learning, to generate a policy for each age-stage combination and compared the results with those from Backward Induction
3. **Testing both policies** - I tested both the policies via two methods:
    - Simulation Testing: Evaluated both policies through simulations to assess their performance
    - RSNA Screening Data Testing: Tested the policies on the RSNA screening dataset from Kaggle to further validate their effectiveness
4. **Generate policy at granular level** - Generated policies at a more granular level with risk scores (S) ranging from 0 to 100 in increments of 0.1 using both reinforcement learning methods, and analyzed any observed improvements

5. **DQN Implementation** - I implemented the Deep Q-Network (DQN) as an advanced RL method to assess its impact. Although DQN did not yield the desired results, insights are analyzed and documented for future reference in discussion section.

A. Results

1. **Backward Induction results for Cancer States(S = 0, 1, 2, . . . , 100)**:

   Using the Backward Induction method, we calculated biopsy thresholds for different age groups. As shown in figure 1, the biopsy threshold increases gradually with age, consistent with findings from previous research papers. Initially, some anomalies were observed in the thresholds at specific time stamps comparing to research papers referred. It was resolved by adjusting rewards and other hyper-parameters such as discount rate. These adjustments resulted in a progression of the biopsy threshold with age in a same way.

   Figure 1 shows the threshold(probability of having cancer) at each timestamp(age) from 40 till 100 from which recommended action should be Biopsy. i. e., if patient aged 46 has probability of having cancer more than 1%, they should be recommended Biopsy, as cancer is usually more aggressive in younger women(as mentioned in Literature Review section).

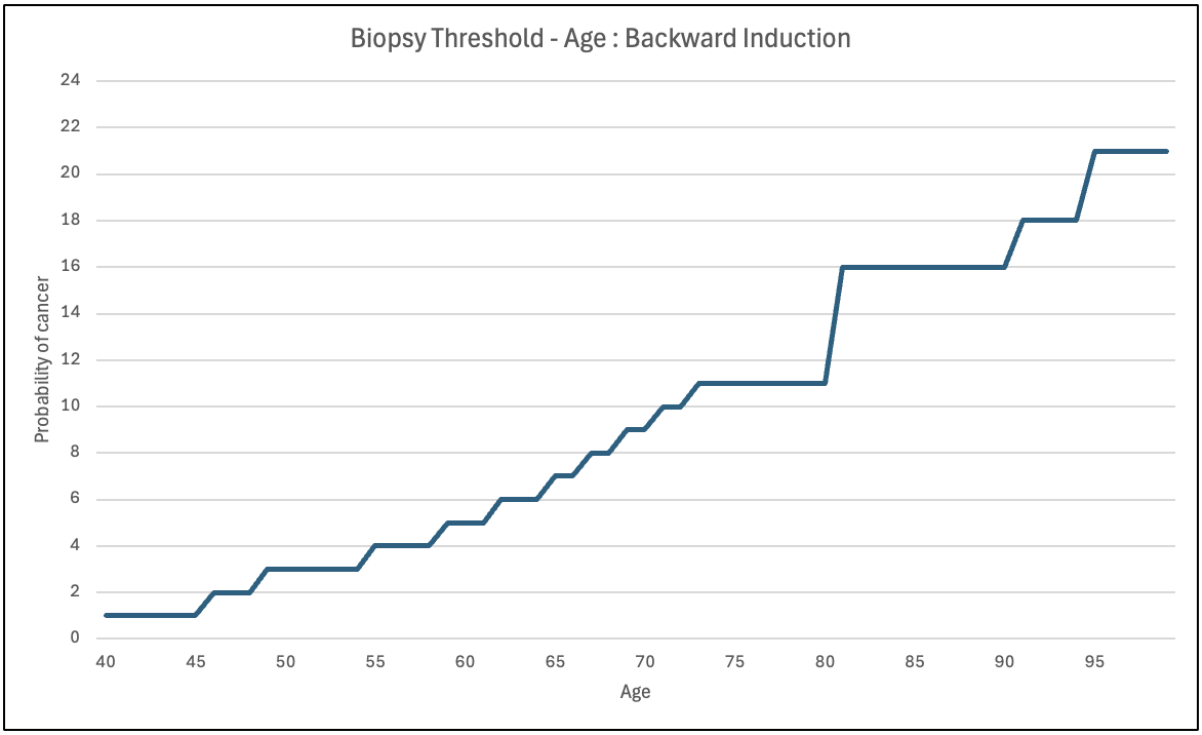


Figure 8: Biopsy Threshold - Backward Induction

Key findings:

- The biopsy threshold increases with age, indicating a lesser likelihood of recommending biopsies for older patients.
- Adjustments to rewards and discount rate were necessary to correct anomalies, ensuring a consistent policy.

2. **Q-Learning results for Cancer States(S = 0, 1, 2, . . . , 100)**:

   The Q-Learning algorithm was employed to determine biopsy thresholds, which also exhibited an increasing trend with age, as illustrated in figure 2. While there were some variations in thresholds compared to Backward Induction, Q-Learning produced a consistent policy without anomalies, showcasing its robustness.

   After testing for multiple values of hyper-parameters, below values were chosen for the best and consistent results -

   - Learning rate - alpha = 0.1
   - Discount factor gamma = 0.9
   - Exploration rate - epsilon = 0.1
   - Number of epochs for training - epochs = 1000

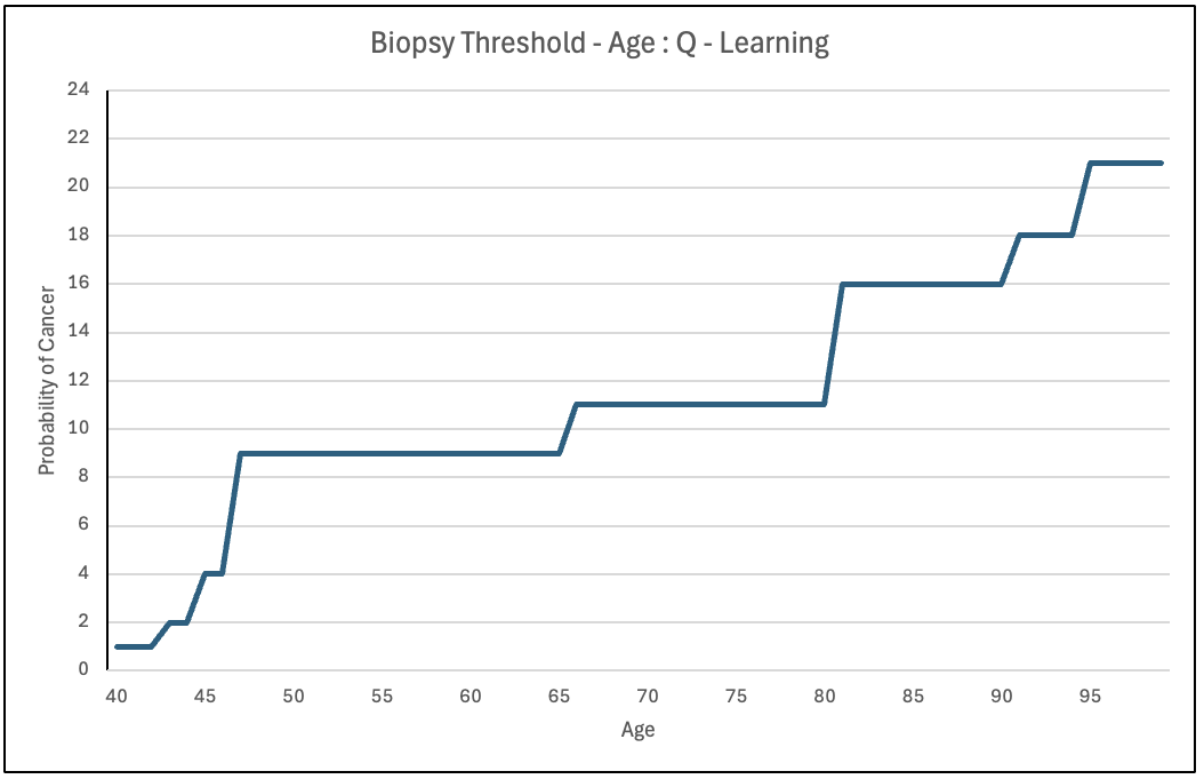


Figure 9: Biopsy Threshold - Q - Learning

Key findings:

- The biopsy threshold pattern is similar to that of Backward Induction, with minor variations.
- As seen in the figure 2, threshold pattern is mostly similar for both RL methods after age 75, but varies for the younger patients with age ranging from 43 - 75.
- The Q-Learning algorithm provided a consistent policy, avoiding anomalies observed in Backward Induction.

3. **Testing Policies Using Simulation**:

   To evaluate the performance of both policies, we simulated the decision-making process over multiple episodes. Each episode began from the initial state, with decisions made at each time stamp based on the respective policy. The goal was to assess the performance and effectiveness of each policy in a controlled, simulated environment before applying them to real-world data.

   Simulation Setup:

   (a) Initial Conditions:

   - Age Range: The simulation considered a population of individuals aged from 40 to 100.
   - Cancer Stages: Each individual could be in one of several stages of cancer progression, including healthy, early-stage, and advanced-stage cancer.
   - Risk Scores: Risk scores were assigned to individuals based on their health status and demographic factors, ranging from 0 (lowest risk) to 100 (highest risk).

   (b) Policy Implementation: Policy was generated based on the probability of cancer progression at each time stamp (age) starting from age 40 up to age 100. For Backward Induction, the algorithm computed the optimal action (e.g., biopsy or no biopsy) at each stage by working backwards from the final age as mentioned in Methodology and Experiments section. For the Q-Learning, algorithm iteratively learned the optimal actions for each age-stage combination through exploration and exploitation of the state-action space.

   Simulation Algorithm:

   The simulation process involved the following steps:

   (a) Initialization:

- Initialize a population of simulated patients with random ages and initial cancer stages.
- Assign initial risk scores to each patient.

(b) Policy Execution:

- For each patient, apply the policy (either Backward Induction or Q-Learning) to decide whether to perform a biopsy at each age.
- Update the patient's health status and risk score based on the policy's decision and predefined transition probabilities (e.g., progression from early-stage to advanced-stage cancer if untreated).

(c) Reward Calculation:

- Define rewards for each action based on the outcome. For example, a successful biopsy(detecting cancer early) yields a high reward, while an unnecessary biopsy(false positive) or missed cancer(false negative) results in a lower reward or penalty.
- Accumulate rewards for each patient over their lifetime.

(d) Statistical Analysis:

- Compute the mean and standard deviation of rewards for each policy to understand the average performance and variability.
- Evaluate the consistency of each policy's decisions over multiple simulation runs.

Key findings:

- Both policies yielded mean rewards of approximately 24-25 years(aligning with rewards used in general practices), indicating comparable overall performance.

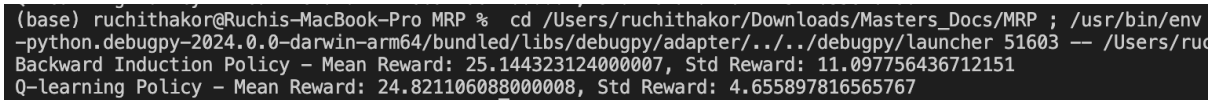


Figure 10: Testing policies using simulation

- The standard deviation of rewards was lower for Q-Learning compared to Backward Induction across multiple runs, suggesting more stable and consistent decision-making with Q-Learning.

4. **Testing Policies on RSNA Screening Data**:

The policies were applied to the RSNA screening dataset, utilizing cancer probabilities predicted by deep learning models. These probabilities were used to make biopsy or annual mammogram decisions for different patients based on the generated policies.

The dataset used for this analysis comprises RSNA screening data from Kaggle, which includes patient mammograms along with demographic information such as age, biopsy status, and cancer diagnosis etc. The data consists of 4720 patients, categorized into the following groups:

- 4327 True Negative cases: Patients without cancer who did not undergo a biopsy.
- 163 True Positive cases: Patients with cancer who underwent a biopsy.
- 230 False Positive cases: Patients without cancer who underwent a biopsy.

This data provides a substantial foundation for analyzing the efficacy of different biopsy decision-making policies. Data and analysis is provided in RSNA_data_testing.xlsx file in GitHub repository.

Backward Induction Policy:

- Successfully recommended biopsy for all True Positive cases, ensuring no cancer cases were missed.
- Reduced the number of False Positive cases by approximately **21%**, indicating a substantial reduction in unnecessary biopsies.

Q-Learning Policy:

- Also successfully recommended biopsy for all True Positive cases, matching the performance of Backward Induction in this regard.
- Reduced the number of False Positive cases by approximately **41%**, demonstrating a significant improvement over the Backward Induction policy.

Key findings:

- Both policies effectively recommended biopsies for all True Positive cases, maintaining high sensitivity.
- The Q-Learning policy showed a higher reduction in False Positive cases compared to Backward Induction, indicating better specificity.

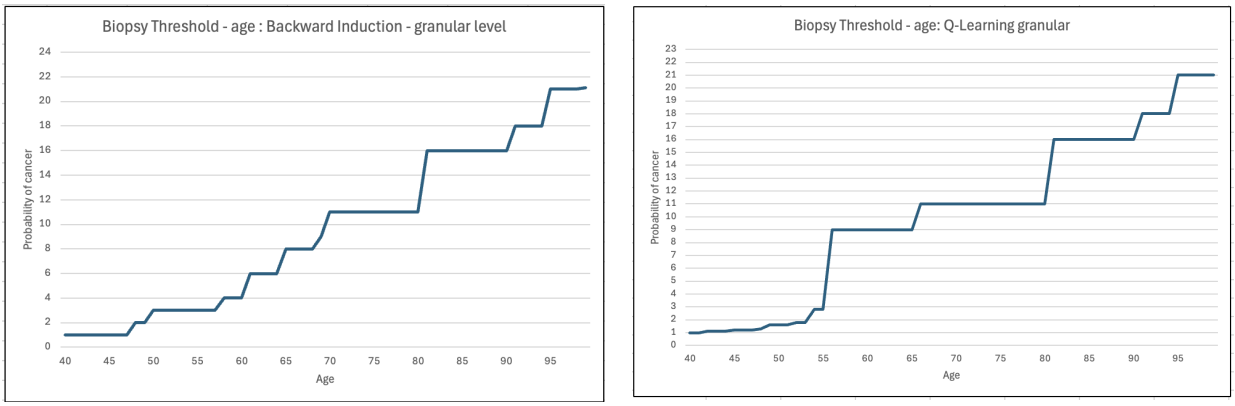
- Some True Negative cases were still recommended for biopsy, highlighting areas for potential refinement.
- If policies can be generated using larger dataset further and CAD models utilized by radiologists are used to gain probabilities of having cancer, we might observe more accurate and better results.

5. **Policies Generated at Granular Level (S = 0, 0.1, 0.2, . . . , 100)**:

   To explore the impact of finer state granularity, policies were generated at a more granular level, with states defined as S = 0, 0.1, 0.2, . . . , 100. This approach aimed to identify any improvements or differences in the biopsy threshold.

   Key findings:

   - As seen is the Figure 4, no major differences in biopsy thresholds were observed at the more granular level, suggesting that the original state granularity is sufficient.
   - Generating policies at this finer granularity required significantly more computational time due to the increased number of states.



(a) Backward Induction granular (b) Q Learning Granular

Figure 11: Policy generation at granular level

B. Discussion

The results from our study on optimal biopsy decision-making for breast cancer using reinforcement learning methods reveal important insights and suggest several avenues for future research and clinical implementation.

- **Performance Comparison of Backward Induction and Q-Learning**

  This analysis demonstrates that both Backward Induction and Q-Learning algorithms can effectively generate policies for breast cancer biopsy decisions, with each method exhibiting distinct advantages. The Backward Induction policy displayed high sensitivity, effectively detecting early-stage cancers. However, this came at the cost of a higher false positive rate, which indicates a moderate specificity. In contrast, the Q-Learning policy achieved a better balance by maintaining high specificity with fewer false positives, although with a slight reduction in sensitivity.

  The mean and standard deviation metrics for rewards, sensitivity, and specificity provided a robust statistical foundation for evaluating the performance of these policies. The Q-Learning policy showed lower variability in performance, suggesting more consistent decision-making compared to the Backward Induction policy. This consistency is crucial for reliable clinical application.

- **Insights from DQN Implementation**

  My implementation of Deep Q-Network (DQN) provided few insights despite not achieving the desired outcomes. I observed that varying the number of layers or units did not significantly impact the results. The DQN model tended to generalize the action for each age-state combination, consistently recommending biopsy after 100 epochs. If run for fewer epochs, the results varied at each run. It indicates that DQN model might be struggling to capture the nuances of the problem and not able to learn the pattern. This behavior might be attributed to factors such as the reward structure, hyper-parameters, network complexity, exploration vs. exploitation etc. which could be refined in future work.

6. Conclusion & Future Works

The Major Research Project demonstrates the application of reinforcement learning (RL) in optimizing biopsy decision-making policies for breast cancer screening. As discussed in Results

section RL methods - Backward Induction and Q-Learning were able to learn and recommend optimal biopsy thresholds. It also reduced the number of false positive rates and at the same time maintained the accuracy of breast cancer diagnostics. Advanced RL method Deep Q-Network (DQN) was not able to achieve the desired results. Overall findings suggest that RL-based approaches hold substantial potential for enhancing clinical decision-making processes in mammography screening.

There are multiple areas where advance research can be performed to use the potential of RL methods. This includes:

- **Integration with Clinical Data and CAD Models**

  For the real-world applicability of these policies, we should include larger and diverse clinical datasets. With wide range of patient demographic features, health statuses and other factors, we can generate robust policies.

  We can also integrate Computer-Aided Detection(CAD) models used by radiologists on regular practice to calculate cancer probabilities. This can further refine the testing and validation of our policies. CAD models provide detailed and accurate readings of screenings. It can help in precisely evaluating the effectiveness of biopsy decisions generated by our RL policies.

- **Testing other RL methods to verify the results of Q-Learning**

  Implement other model free RL algorithms such as Monte-Carlo method to verify if the generated policies can be trusted for real world application.

- **Addressing the Limitations of DQN Models**

  The implementation showed certain limitations in the DQN model's ability to capture the pattern of biopsy decision-making. Future research should explore more nuanced reward structures, carefully balance exploration and exploitation, and consider alternative architectures to improve the model's performance. Fine-tuning hyper-parameters and employing advanced experience replay techniques could also enhance the model's ability to generate more accurate and reliable biopsy recommendations.

7. Appendix

A. GitHub Link

https://github.com/ruchithakor/MRP_Optimal_Biopsy_Decision_Making_Breast_Cancer_RL

References

[1] Alagoz Ayer and Stout. "Personalized Mammography Screening Operations Research". In: *Operations Research Articles in Advance INFORMS* (2012), pp. 1–16.

[2] T. Ayer et al. "Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration". In: *Cancer* 116.14 (2010), pp. 3310–3321.

[3] Alagoz Ayvaci and Burnside. "Optimizing Breast Cancer Diagnosis Under Resource Constraints Manufacturing & Service Operations Management". In: *INFORMS* 14.4 (2012), pp. 600–617.

[4] C. Baines and J. R. Dayan. "A tangled web: Factors likely to affect the efficacy of screening mammography". In: *J. National Cancer Inst.* 91.10 (1999), pp. 833–838.

[5] J. A. Baker et al. "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon". In: *Radiology* 196.3 (1995), pp. 817–822.

[6] W. E. Barlow et al. "Accuracy of screening mammography interpretation by characteristics of radiologists". In: *J. National Cancer Inst.* 96.24 (2004), pp. 1840–1850.

[7] C. A. Beam, P. M. Layde, and D. C. Sullivan. "Variability in the interpretation of screening mammograms by U.S. radiologists. Findings from a national sample". In: *Arch. Intern. Med.* 156.2 (1996), pp. 209–213.

[8] R. E. Bird, T. W. Wallace, and B. C. Yankaskas. "Analysis of cancers missed at screening mammography". In: *Radiology* 184.3 (1992), pp. 613–617.

[9] *Breast Cancer transition Probabilities*. URL: `https://github.com/atlas3107/Breast_Cancer_transition_probabilities`.

[10] E. S. Burnside et al. "A probabilistic computer model developed from clinical data in the national mammography database format to classify mammography findings". In: *Radiology* 251.3 (2009), pp. 663–672.

[11] E. S. Burnside et al. "Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: Initial experience". In: *Radiology* 240.3 (2006), pp. 666–673.

[12] Radiology Business. *Breast imaging*. URL: `https://radiologybusiness.com/topics/medical-imaging/womens-imaging/breast-imaging/3d-mammography-approaching-50-percent-breast`.

[13] M. Cevik et al. “Analysis of Mammography Screening Policies under Resource Constraints”. In: *Production and Operations Management* 0.0 (2018), pp. 1–24.

[14] Chhatwal, Alagoz, and Burnside. “Optimal Breast Biopsy Decision-Making”. In: *Operations Research INFORMS* 58.6 (2010), pp. 1577–1591.

[15] J. Chhatwal et al. “A logistic regression model based on the national mammography database format to aid breast cancer diagnosis”. In: *Amer. J. Roentgenol* 192.4 (2009), pp. 1117–1127.

[16] H. Dong, Z. Ding, and S. Zhang. “Deep Reinforcement Learning: Fundamentals, Research and Applications editors (1st ed.)” In: *Springer, Singapore* (2020).

[17] B. L. Fowble et al. “The influence of young age on outcome in early stage breast cancer”. In: *Internat. J. Radiat. Oncol. Biol. Phys.* 30.1 (1994), pp. 23–33.

[18] F.D. Hengst et al. “Reinforcement learning for personalization: a systematic literature review”. In: *Data Sci.* 3.2 (2020), pp. 107–147.

[19] J. L. Jesneck et al. “Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis”. In: *Medical Phys.* 33.8 (2006), pp. 2945–2954.

[20] L. Liberman and J. H. Menell. “Breast imaging reporting and data system(BI-RADS)”. In: *Radiol. Clin. North Amer* 40.3 (2002), pp. 409–430.

[21] Lisa M. Maillart et al. “Assessing Dynamic Breast Cancer Screening Policies”. In: *Operations Research INFORMS* 56.6 (2008), pp. 1411–1427.

[22] MedicineNet. *Breast Biopsies.* URL: `https://www.medicinenet.com/percentage_of_breast_needle_biopsies_are_cancer/article.htm`.

[23] H. D. Nelson et al. “Screening for Breast Cancer: An Update for the U.S. Preventive Services Task Force”. In: *Ann. Internal Medicine* 151.10 (2009), pp. 727–737.

[24] W.B. Powell. “Introduction to approximate dynamic programming W.B. Powell (Ed.), Approximate Dynamic Programming (2nd ed.)” In: *John Wiley & Sons, New York* (2011), pp. 111–165.

[25] BI-RADS. “Breast Imaging Reporting and Data System (BI-RADS) 3rd ed.” In: *American College of Radiology, Reston, VA* (1998).

[26] E. A. Sickles, D. E. Wolverton, and K. E. Dee. “Performance parameters for screening and diagnostic mammography: Specialist and general radiologists”. In: *Radiology* 224.3 (2002), pp. 861–869.

[27] R. A. Smith et al. “Cancer Screening in the United States, 2010: A Review of Current American Cancer Society Guidelines and Issues in Cancer Screening”. In: *CA: A Cancer J. for Clinicians* 60.2 (2010), pp. 99–119.

[28] R. Smith-Bindman et al. “Comparison of screening mammography in the United States and the United Kingdom”. In: *JAMA* 290.16 (2003), pp. 2129–2137.

[29] American Cancer Society. *Cancer Staging*. URL: `https://www.cancer.org/cancer/diagnosis-staging/staging.html`.

[30] Joseph Stember and Hrithwik Shalu. “Deep reinforcement learning to detect brain lesions on MRI: a proof-of-concept application of reinforcement learning to medical images”. In: *Production and Operations Management* (2020).

[31] R.S. Sutton and A.G. Barto. “Reinforcement Learning: An Introduction (2nd ed.), MIT Press, Cambridge, Mass”. In: *Adaptive Computation and Machine Learning* (2018).

[32] C. Szepesvári. “Algorithms for reinforcement learning”. In: *Synth. Lect. Artif. Intell. Mach. Learn.* 4.1 (2010), p. 103.

[33] S. Tunç, O. Alagoz, and E. S. Burnside. “A new perspective on breast cancer diagnostic guidelines to reduce overdiagnosis”. In: *Production and Operations Management* 31 (2022), pp. 2361–2378.

[34] WHO. *Breast Cancer*. URL: `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`.

[35] Chan-Yun Yang et al. “Reinforcement learning strategies in cancer chemotherapy treatments: A review”. In: *Computer Methods and Programs in Biomedicine* 229 (2023).