Received: 7 September 2023 | Revised: 20 March 2024 | Accepted: 14 April 2024

DOI: 10.1002/mp.17100

RESEARCH ARTICLE

MEDICAL PHYSICS

Clinical VMAT machine parameter optimization for localized prostate cancer using deep reinforcement learning

William T. Hrinivich | **Mahasweta Bhattacharya** | **Lina Mekki** | **Todd McNutt** | **Xun Jia** | **Heng Li** | **Daniel Y. Song** | **Junghoon Lee**

Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, Maryland, USA

Correspondence
William T. Hrinivich, Johns Hopkins Proton Therapy Center, 5255 Loughboro Rd NW, Washington, DC 20016, USA.
Email: whriniv1@jhmi.edu

Funding information
Commonwealth Fund

Abstract

Background: Volumetric modulated arc therapy (VMAT) machine parameter optimization (MPO) remains computationally expensive and sensitive to input dose objectives creating challenges for manual and automatic planning. Reinforcement learning (RL) involves machine learning through extensive trial-and-error, demonstrating performance exceeding humans, and existing algorithms in several domains.

Purpose: To develop and evaluate an RL approach for VMAT MPO for localized prostate cancer to rapidly and automatically generate deliverable VMAT plans for a clinical linear accelerator (linac) and compare resultant dosimetry to clinical plans.

Methods: We extended our previous RL approach to enable VMAT MPO of a 3D beam model for a clinical linac through a policy network. It accepts an input state describing the current control point and predicts continuous machine parameters for the next control point, which are used to update the input state, repeating until plan termination. RL training was conducted to minimize a dose-based cost function for prescription of 60 Gy in 20 fractions using CT scans and contours from 136 retrospective localized prostate cancer patients, 20 of which had existing plans used to initialize training. Data augmentation was employed to mitigate over-fitting, and parameter exploration was achieved using Gaussian perturbations. Following training, RL VMAT was applied to an independent cohort of 15 patients, and the resultant dosimetry was compared to clinical plans. We also combined the RL approach with our clinical treatment planning system (TPS) to automate final plan refinement, and creating the potential for manual review and edits as required for clinical use.

Results: RL training was conducted for 5000 iterations, producing 40 000 plans during exploration. Mean ± SD execution time to produce deliverable VMAT plans in the test cohort was 3.3 ± 0.5 s which were automatically refined in the TPS taking an additional 77.4 ± 5.8 s. When normalized to provide equivalent target coverage, the RL+TPS plans provided a similar mean ± SD overall maximum dose of 63.2 ± 0.6 Gy and a lower mean rectum dose of 17.4 ± 7.4 compared to 63.9 ± 1.5 Gy ($p = 0.061$) and 21.0 ± 6.0 ($p = 0.024$) for the clinical plans.

Conclusions: An approach for VMAT MPO using RL for a clinical linac model was developed and applied to automatically generate deliverable plans for localized prostate cancer patients, and when combined with the clinical TPS shows potential to rapidly generate high-quality plans. The RL VMAT approach shows promise to discover advanced linac control policies through trial-and-error, and algorithm limitations and future directions are identified and discussed.

KEYWORDS
artificial intelligence, automation, deep learning, prostate cancer, reinforcement learning, VMAT

1 | INTRODUCTION

Volumetric modulated arc therapy (VMAT) is a clinically standard radiotherapy technique for the treatment of prostate and many other cancers, providing fast treatment delivery and highly conformal dose distributions by continuously delivering radiation during gantry rotation.[1] VMAT plan optimization requires the definition of a sequence of multi-leaf collimator (MLC) positions and monitor units (MU) at discretized gantry angles over the range of the arc, and remains a complex and time-consuming process despite many advancements toward efficiency improvements and automation. A single plan may incorporate $>10^4$ individual continuous machine parameter values, which must satisfy dose objectives and the physical limitations of the linear accelerator (linac) including leaf translation, gantry rotation, and dose rate limits. VMAT machine parameter optimization (MPO) represents a high-dimensional non-convex problem, and a variety of numerical inverse VMAT MPO approaches have been developed as reviewed by Unkelbach et al.[2] These approaches involve defining a cost function based on dose objectives and weights, which is minimized with respect to the machine parameters through arc sequencing methods, direct aperture optimization, and leaf trajectory optimization among other recent approaches.[3] These algorithms enable the generation of high-quality VMAT plans within clinical time constraints, including GPU-based column generation algorithms with execution times of 18–55 s,[4,5] but have common limitations including high sensitivity to input dose objectives and weights used to define the cost function. The interaction between input dose objectives and resultant treatment plan is highly complex and non-linear depending on patient-specific geometry, making definition of patient specific dose objectives a challenging clinical problem. In general, this complex behavior is overcome by human planners who manually edit dose objectives for each treatment plan in a trial-and-error fashion until an acceptable plan is created. These characteristics of existing VMAT MPO algorithms are major contributors to clinical challenges of radiotherapy planning time and variability.

Many approaches have been investigated to overcome these clinical challenges in VMAT planning through automation, and typically focus on automatic generation of input dose objectives while using the same VMAT MPO algorithms as used for manual planning.[6] These approaches include supervised machine learning (ML) algorithms such as knowledge-based planning (KBP)[7,8] and dose synthesis.[9–12] These ML approaches are inherently designed to generate dose objectives leading to plans matching dosimetry of those in the training set, placing an upper bound on expected plan quality and critical dependence on plans in the training set. Fluence-map prediction approaches have been developed for intensity-modulated radiotherapy (IMRT) replacing a major component of MPO,[13–15] but to our knowledge have also not been applied to VMAT and again represent a supervised learning approach.

Recently, deep reinforcement learning (RL) has been investigated for automated radiotherapy planning, which involves training an algorithmic agent to control a dynamic system through trial-and-error.[16] By incrementally learning from successes and mistakes over many iterations, deep RL may learn control strategies that not only match but exceed the performance of existing approaches. Super-human RL performance has been demonstrated in multiple problem domains including board games,[17] video games,[18,19] and was recently used to fine-tune large language models including Chat GPT-4.[20,21] For radiotherapy planning, RL has also been applied to the prediction and refinement of dose objectives for multiple inverse planning applications, mimicking human planners and depending on existing numerical MPO approaches. These virtual treatment planner network (VTPN) applications include high-dose-rate brachytherapy for cervical cancer[22] and IMRT for localized prostate cancer,[23,24] but to our knowledge have not been applied to VMAT due in part to high computational cost of existing VMAT MPO which makes training prohibitively time-consuming. While demonstrating high performance, the VTPN approach would also benefit from improvements in VMAT MPO.

Our group previously proposed applying deep RL directly to VMAT machine parameter control, thereby replacing both the cost function tuning and existing VMAT MPO components by an algorithmic agent which directly generates machine parameter values.[25] By giving the agent direct control of the machine parameters, it may be possible to discover linac control policies that exceed the performance of existing algorithms. VMAT MPO using RL could also significantly reduce computational cost and mitigate iterative plan computation and parameter refinement required by existing approaches, reducing time required to produce deliverable treatment plans. Our previous study demonstrated proof-of-principle of VMAT MPO using RL using a deep-Q learning approach[19] and simplified 2D beam model for localized prostate cancer. While useful for research and experimentation, this approach could not be applied clinically due to the simple 2D nature and

limitations of the deep-Q approach, including the control of machine parameters in discretized steps.

The purpose of the current study is to describe and validate an RL-based VMAT MPO approach for localized prostate cancer for a full clinical beam model, referred to herein as "RL VMAT," and compare it to clinical plans produced by human planners in a commercial treatment planning system (TPS). Extending our previous RL VMAT approach from 2D to 3D required several technical developments including interfacing a 3D collapsed cone convolution (CCC) superposition photon dose engine with open-source deep learning libraries, replacing deep-Q RL with a policy-gradient RL approach enabling prediction of continuous machine parameters,[26] increasing the amount of training data, and implementing data augmentation during training to mitigate over-fitting. We also integrated the RL VMAT approach with our clinical TPS to demonstrate a workflow which automatically produces a deliverable VMAT plan matching or exceeding clinical dose objectives, but retains the ability for human planners to fine-tune the plan through inverse optimization if needed.

2 | METHODS

2.1 | Beam model

We developed a GPU-based 3D CCC megavoltage photon dose calculation algorithm which was compiled as a Python package, enabling simple integration with open-source deep learning packages as required for RL algorithm development. The dose engine was modeled to match the 6 MV beam energy of our Axesse linac (Elekta, Stockholm, Sweden). The linac has an Agility MLC with opposing banks of 80 leaves with 5 mm width at isocenter, maximum individual leaf speed of 3.5 cm/s, additional carriage speed of 3.0 cm/s and allows for interdigitation.[27] The dose engine used concepts developed by Jacques et al.[28] with simplifications to streamline beam modeling and decrease execution time, focusing on characteristics critical for RL training. Specifically, we chose the kernel parameterization presented by Ahnesjo including the 6 MV kernel fitting parameters.[29] Further beam modeling details are provided in Appendix A.

2.2 | Policy network for linac control

VMAT plans are typically parameterized through a set of control points corresponding to N discrete, uniformly spaced gantry angles; however, MLC positions and dose rates at each control point vary continuously. We chose a policy network approach, which complements this VMAT parameterization by controlling dynamic systems through state-action pairs in which the current environment state s_t is taken as input, a set of continuous action values a_t are predicted as output according to

$$a_t = \pi_\theta\ (s_t) \tag{1}$$

where π_θ represents the policy function, approximated by a neural network with weights θ. The state s_t is then updated by a_t through the operator $s_{t+1} = f(s_t, a_t)$. This operation leads to a new state, which is again taken as input by the policy network. This process is repeated for N steps until process termination, or final control point.

For 3D VMAT control, we designed state s_t with two components including a multi-channel 3D array and a 1D array designed to represent the state of the patient and treatment plan at the current control point as follows. The 3D array had three channels including (1) a label map containing the planning target volume (PTV), (2) a label map containing the organs at risk (OARs) including external body contour, rectum, bladder, and femoral heads, and (3) the cumulative dose grid at the current control point. The PTV and OARs were represented in separate channels to accommodate overlap of the PTV and OARs. These 3D arrays were centered at isocenter with $64 \times 64 \times 64$ dimensions and uniform 5 mm voxel spacing. The 1D array consisted of 62 floating point values representing the central 60 MLC leaf positions, dose rate, and gantry angle of the current control point. We designed the action a_t as 61 floating point values including the central 60 MLC leaf positions (30 leaf pairs) and dose rate of the control point at the next gantry angle. Limiting MLC control to the central 30 leaf pairs reduced network complexity while still covering the target for the patients in this study. Non-controlled leaves were closed, and the collimator angle, couch angle, gantry start angle, and gantry stop angle were fixed.

The operator $s_{t+1} = f(s_t, a_t)$ computes the control point defined by a_t at gantry angle corresponding to s_{t+1}, and adds it to the dose array and updates the 1D array of machine parameters. The operator $f(s_t, a_t)$ also enforced leaf travel limits from one control point to another, thereby ensuring deliverability of the final plan. The VMAT plan consists of the list $\boldsymbol{a}$ of N actions. In our previous RL version, the deep-Q network controlled individual pairs of leaves based on a 2D slice of the dose distribution aligned with those leaves, which was dynamically re-sampled during training.[25] In our current RL version, the policy network simultaneously controls 30 pairs of leaves and does not require dynamic dose grid resampling during training or execution.

The policy function π_θ was implemented as a convolutional neural network (CNN) with three convolutional layers and three fully connected layers as shown in Figure 1a, modified from the deep-Q network used in our previous study.[25] The multi-channel 3D array component of s_t is taken as input 1 in the first convolutional layer. Following the third convolutional layer, the network is flattened and the 1D state array is concatenated with

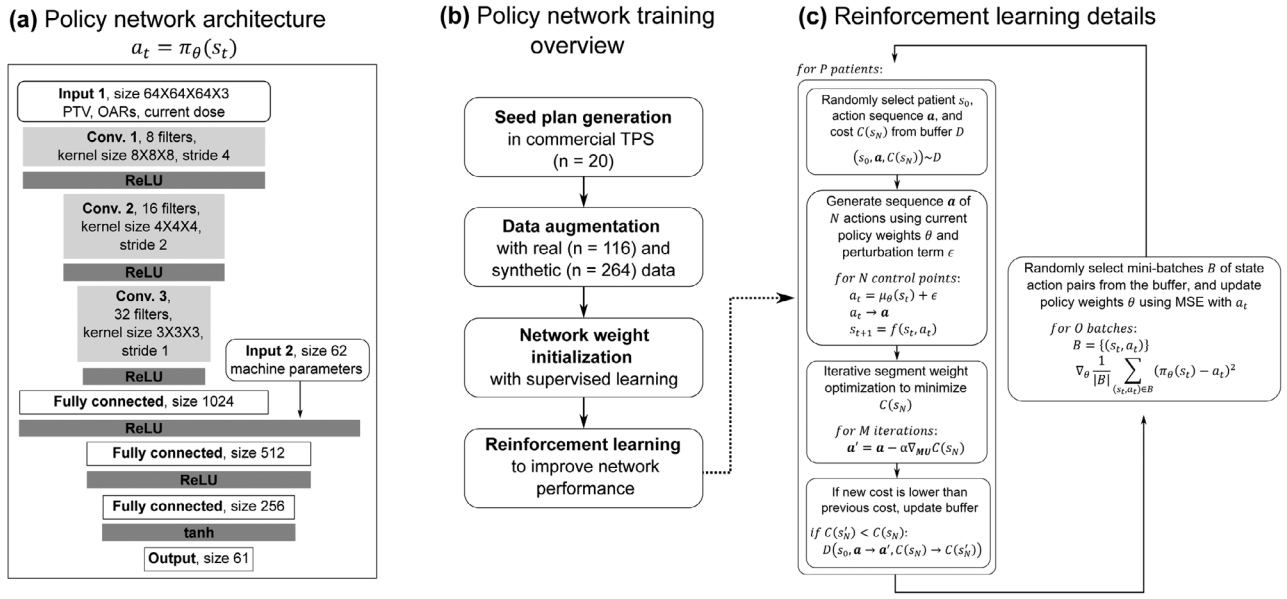

FIGURE 1 (a) Schematic indicating the policy network design. (b) Flowchart of major steps of the policy network training algorithm. (c) Flowchart indicating details of the RL training component.

the layer, denoted as input 2. We employed ReLU activation functions between all layers except for the final layer, which employed a tanh activation function providing output action values varying continuously from −1.0 to 1.0. These actions were then converted to MLC positions by multiplying the 60 MLC actions by 75.0 mm, and computing the MU based on the single MU action $a_{t,MU}$ using

$$MU = 20.0 * (a_{t,MU} + 1.1) \quad (2)$$

These action conversions inherently limited leaf travel limits to the range [−75.0 mm, 75.0 mm] and the MU per control point to the range [2.0 MU, 42.0 MU], which were found to be sufficient for the VMAT plans in this study but would require modification for larger targets.

2.3 | Policy network training overview

The major steps of policy network training are provided in Figure 1b. Given the increase in complexity of the 3D VMAT control problem compared to our previous 2D VMAT control approach, we employed a network initialization approach using supervised learning, analogous to approaches used for other sparse large scale RL problems such as real-time strategy-based game control.[18] We then employed RL on a larger set of training data to improve network performance.

2.3.1 | Data

CT scans and contours from 186 localized prostate cancer patients previously treated at our institution were anonymized and used in this study through a retrospective IRB approved protocol. These cases were randomly split into independent training (n = 136), validation (n = 35), and test (n = 15) cohorts. For the test cohort patients, we also extracted the clinical VMAT treatment plans produced by human planners through our standard clinical workflow which were reviewed and approved for treatment by the attending physician.

2.3.2 | Seed plan generation

A subset of 20 training cases were imported into our clinical TPS (RayStation 2023B, RaySearch, Stockholm, Sweden). Single arc VMAT plans were produced for each patient with 4° gantry angle spacing and 15° collimator angle with prescription dose of 60 Gy in 20 fractions. We employed our institutional dose objectives when optimizing the plans, derived from the CHHiP protocol.[30] During seed plan optimization, the planner could use any combination of dose objectives and any number of iterations or objective edits to achieve an acceptable final plan, similar to clinical planning. The resultant 20 seed plans were then exported from the TPS to initialize training.

TABLE 1 Dose objective values and weights used to construct the default cost function.

Structure	Metric	Value	Weight
PTV	Max. dose	62.4 Gy	100
PTV	V60 Gy	100%	50
PTV	Min. dose	61.2 Gy	20
Ring 1 (0.1–1.1 cm from PTV)	Max. dose	58.2 Gy	1
Ring 2 (1.2–2.2 cm from PTV)	Max. dose	42.0 Gy	0.5
Ring 3 (2.3–4.3 cm from PTV)	Max. dose	36.0 Gy	0.1
Ring 4 (4.4–6.4 cm from PTV)	Max. dose	27.0 Gy	0.1
Rectum	Max. dose	61.8 Gy	1.0
Rectum	Mean dose	0.0 Gy	5.0
Bladder	Max. dose	61.8 Gy	1.0
R femoral head	Max. dose	30.0 Gy	1.0
R femoral head	Mean dose	0.0 Gy	0.2
L femoral head	Max. dose	30.0 Gy	1.0
L femoral head	Mean dose	0.0 Gy	0.2

2.3.3 | Data augmentation

Given the complexity of the policy network and number of iterations required for model convergence, we further augmented the 136 training cases by randomly selecting patients with replacement, applying random scaling and 3D rotations to the CT, and adding the result back into the training cohort. Scaling and 3D rotation factors were sampled from uniform distributions from [0.7, 1.3] and $[-3°, 3°]$, respectively. This process was repeated to generate 264 additional "synthetic" training cases, leading to 400 training cases in total.

In order to use all 400 training cases for policy weight initialization, the 20 initial seed plans were initially mapped to the additional 380 training cases as follows. For each of the 380 additional cases, each of the 20 seed plans were transferred via rigid registration and recomputed on the CT associated with the additional case. The cost associated with each plan was computed based on a set of default dose volume histogram (DVH) objectives. Specifically, the default cost C was computed based on the mean squared error (MSE) between a set of PTV and OAR dose metrics d_i and corresponding objectives o_i with associated weights w_i

$$C = \sum_{i \in S} w_i [\max(d_i - o_i, 0)]^2 \quad (3)$$

For minimum dose objectives on targets, d_i and o_i were multiplied by -1 to satisfy the max() operator. Specific dose objectives and weights used in this study are provided in Table 1. The seed plan resulting in minimum cost for each case was selected, and further refined through gradient descent using the default cost function for 100 iterations. This process resulted in a sub-optimal, but reasonable VMAT plan for each training case which could be used to initialize the network weights through supervised learning.

2.3.4 | Policy weight initialization with supervised learning

The initial VMAT plans were used to initialize the policy network weights as follows. Each plan was re-computed to generate the state-action pair (s_t, a_t) for each control point, where actions were taken directly from the initial VMAT plans. The cost associated with the terminal state $C(s_N)$ was computed, and all state-action pairs and final cost were added to the buffer D. The policy network was then trained to predict the actions based on the input states by sampling mini-batches from D and minimizing a MSE loss function. We used the Adam optimizer[31] with learning rate of 2×10^{-4}, beta of 0.99, mini-batch size of 64, and 10^3 iterations for network initialization.

2.3.5 | Exploration using Gaussian perturbations

We employed action perturbations following a Gaussian distribution for parameter exploration during RL training, in which a perturbation term ϵ was added to the action predicted by the policy network during training. In this study, ϵ was randomly generated from a Gaussian distribution with a mean of 0.0 and a standard deviation of 0.01, which remained constant for the duration of training.

2.3.6 | Reinforcement learning

The major components of the RL policy network training are shown in Figure 1c. We employed a deep deterministic policy gradient (DDPG) approach,[32] in which the policy network produces deterministic action predictions based on the input state. We trained the network to minimize a dose-based cost rather than maximize a reward, similar to conventional inverse VMAT optimization.[2] The VMAT control problem is episodic, consisting of a finite number of steps defined by the number of control points N, and plan dosimetry depends on the summation of dose from all N control points. The cost C was then computed based on the terminal state s_N, including the summation of dose from all control points, indicated by $C(s_N)$.

Each iteration of RL training involves exploration to update the buffer followed by policy network training based on the updated buffer. The VMAT RL approach is modified from actor-critic-based policy gradient approaches[26] to take advantage of the fact that the VMAT cost function is differentiable with respect to

the machine parameters, so we can replace the critic network with direct calculation of the cost function and significantly simplify the algorithm. The entire RL algorithm was implemented in Python using PyTorch on a Quad workstation (Lambda Labs, San Francisco CA), with 4 Quadro RTX 5000 GPUs (Nvidia, Santa Clara CA) with 16 GB of memory each, a Ryzen Threadripper 3970× 32-core CPU (AMD, Santa Clara CA), and 256 GB of RAM.

Exploration to update the buffer

An iteration of RL training begins by randomly selecting a single patient in the training set. The corresponding entry in the buffer D is selected including the initial state s_0, the existing action sequence $\boldsymbol{a}$, and existing terminal cost $C(s_N)$. A new action sequence is generated using the current policy network weights θ. This action sequence is then refined using iterative segment weight optimization to minimize the default cost function defined in Equation (3) for M iterations, where M was set to 10 in this study. The terminal cost of the refined action sequence $C(s'_N)$ is then compared to the buffer cost entry $C(s_N)$. If the updated cost is less than the buffer cost, the buffer entries are updated, replacing the action sequence $\boldsymbol{a}$ with $\boldsymbol{a}'$ and the terminal cost $C(s_N)$ with $C(s'_N)$. If the updated cost is not lower than the buffer cost due to exploration perturbations or incomplete network training, then the buffer remains unchanged. To improve exploration efficiency, this process was executed for eight randomly selected patients per iteration of RL training. Plan generation for eight patients was executed in parallel across four GPUs.

Policy network training

Following the buffer update step, the policy network weights θ were updated using the Adam optimizer to minimize a MSE loss on action predictions using mini-batches B of 64 state-action pairs (s_t, a_t) from the buffer D. For each iteration of RL training, the policy networks were updated using O mini-batches, where O was set to 10 in this study. The updated policy network weights are then used for the next iteration of plan exploration. In this way, each iteration of RL training includes the generation of 8 VMAT plans to explore the parameter space and 10 iterations of policy network training.

2.4 | Combining RL and TPS for test plan generation

Although the RL approach can automatically produce VMAT plans on unseen cases following training, a clinically realistic implementation of the approach must enable user modification and further optimization of the final plan if required. To satisfy this requirement, we combined the RL approach with our commercial TPS as shown in Figure 2. In this workflow, the policy network is used to generate the initial VMAT machine parameters followed by 10 iterations of segment weight optimization. The corresponding machine parameters are sent back to the TPS and recomputed with the commercial dose engine with 3 mm dose grid, which is referred to as the RL plan. OAR doses are then evaluated and used to automatically create an RL-based cost function C' in which dose objectives are set based on a combination of protocol objectives and the OAR dose metrics achieved through RL. This RL-based cost function is then used to further optimize the RL plan using gradient descent within the commercial TPS. We refer to this automatically generated result as the RL+TPS plan. Although not performed in this study, in a clinical implementation the RL+TPS plan could be evaluated by the physician, and feedback could be incorporated by manually editing the cost function following a traditional planning workflow.

Details of the terms in the RL-based cost function are provided in Table 2. Since the RL VMAT algorithm provides a high-quality sequence of machine parameters, a small number of simple dose objectives were required for the TPS-based refinement. Objectives and weights for PTV minimum dose, maximum dose, and dose fall-off for overall plan conformity were set as constants based on the clinical protocol. Objectives for the rectum and femoral heads were patient-specific and determined by subtracting constants from the corresponding dose metric derived from the RL plan, denoted d'. The constants were found to benefit the quadratic cost function in the TPS, ensuring that the dose to these structures did not increase through the additional steps of gradient descent. In the case of the femoral heads, plan symmetry was encouraged by selecting the minimum d' value between left and right femoral heads and using the result for both femoral heads.

2.5 | Dosimetric comparison with clinical plans

Following training, the RL and RL+TPS algorithms were applied to an independent cohort of 15 anonymized prostate cancer patients previously treated using VMAT. The retrospective clinical dose distributions were used as benchmarks for the RL-based approaches, and were produced in the Pinnacle TPS (Philips, Madison WI) by clinical dosimetrists through conventional inverse optimization, and reviewed and approved by the attending physician for treatment. Pinnacle was the longstanding clinical TPS at our institution while RL development and clinical commissioning occurred in RayStation. We, therefore, uniformly collected clinical treatment plans from Pinnacle as high-quality benchmarks for the RL algorithm, which was then implemented in RayStation.

Based on the mean cost observed in the validation cohort over the course of training, the final network weights from training iteration 4750 were used for

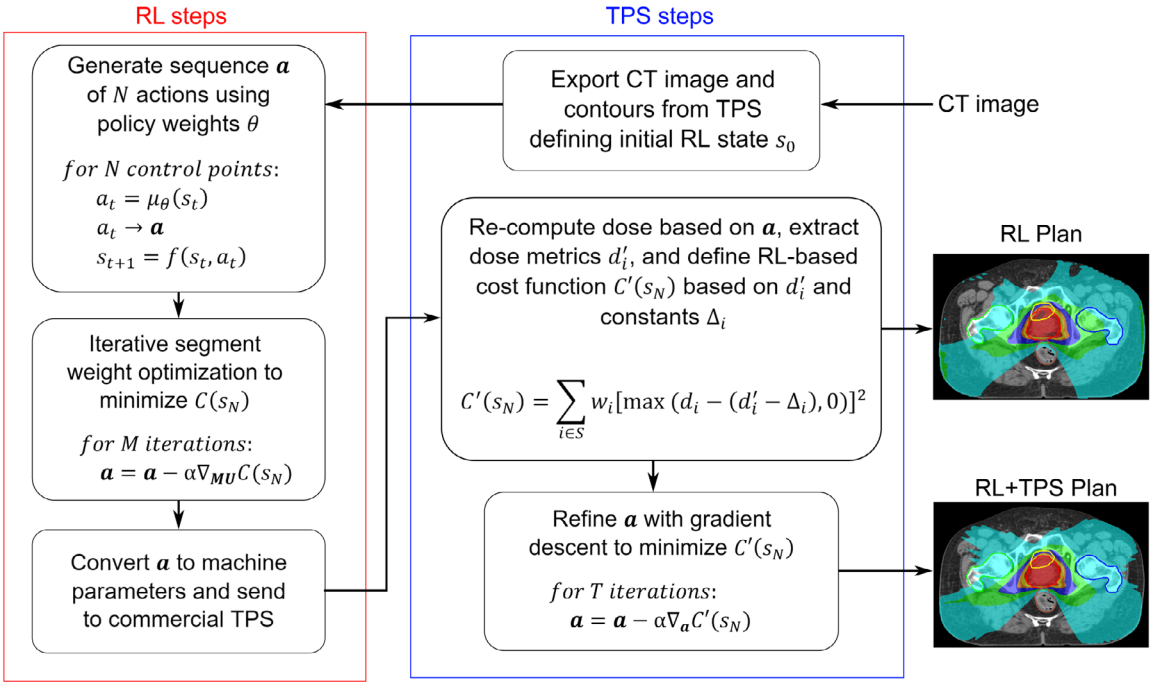

FIGURE 2 Flowchart describing the algorithm used to generate RL and RL+TPS VMAT plans in unseen patients. The portions outlined in blue are completed within the TPS, and portions outlined in red are completed by our RL algorithm outside of the TPS.

TABLE 2 Dose objective values and weights used to create the RL-based cost functions in the RL+TPS approach.

Structure	Metric	Value	Weight
PTV minus rectum	Min. D99%	61.0 Gy	2000
PTV	Min. dose	58.0 Gy	1000
Body	Max. dose	63.0 Gy	5000
Body	Dose fall off	60.0 Gy to 0.0 Gy over 2.0 cm	1
Rectum*	Mean dose	d' – 5.0 Gy	10
Bladder	Max. dose	61.5 Gy	10
R femoral head*	Max. dose	d' – 5.0 Gy	10
L femoral head*	Max. dose	d' – 5.0 Gy	10

The structures with RL-based dose objectives are indicated by asterisks. Symbol d' denotes metrics values derived from the RL plans.

RL inference for the RL and RL+TPS approaches. The RL+TPS approach made use of the automatic cost function definition and 100 iterations of gradient-descent per patient with no further cost function tuning.

The clinical and RL VMAT approaches were dosimetrically compared as follows. The RL+TPS and clinical treatment plans were each scaled to provide equivalent target coverage within each patient such that PTV V60 Gy was between 95% and 98%, thereby allowing comparison of the remaining dose metrics. The RL dose distributions were saved without any scaling due to relatively low target coverage for some patients making simple scaling ineffective. Cumulative DVHs and clinical dose metrics were extracted from the clinical, RL, and RL+TPS plans using our commercial TPS including mean dose (D_{mean}), maximum dose (D_{max}), minimum dose (D_{min}), and Paddick conformity index (CI).[33] Mean DVHs were computed across patients for each planning approach. Mean metric values across patients were compared between the clinical and each RL-based planning approach, and statistical significance of mean differences was assessed using paired *t*-tests.

2.6 | Evaluating policy network produced by supervised learning

To demonstrate the importance of the RL component of training to final policy network performance relative to the supervised learning (SL)-based initialization,

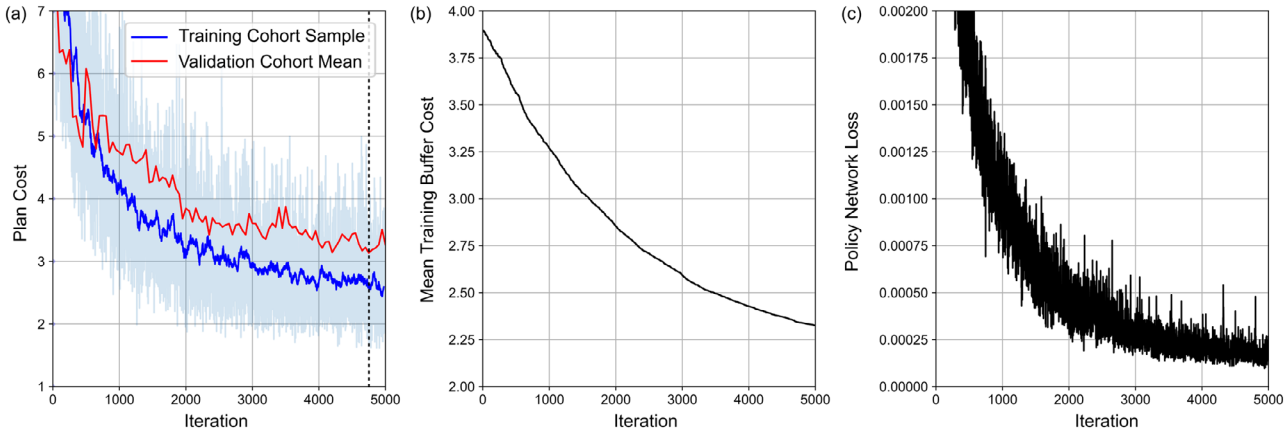

FIGURE 3 (a) Plot of plan cost versus RL training iteration for individual patients sampled from the training cohort in blue, where the solid line is the data smoothed with a uniform 50 iteration window and mean plan cost across the 35 validation patients in red. (b) Plot of mean training buffer cost versus RL training iteration. (c) Plot of policy network loss versus RL training iteration.

the policy network weights generated following the SL portion of training were used to generate VMAT plans in the independent test cohort. Dose metrics associated with these "SL VMAT" plans were compared to the final RL VMAT plans.

3 | RESULTS

3.1 | Training metrics

Network initialization and 5000 RL iterations took 16 h to complete. Plots of training metrics versus iteration in the RL phase are provided in Figure 3, including the training cohort in blue and validation cohort in red. The cost associated with the training cohort shown in light blue displays noise associated with parameter exploration so it is also displayed with smoothing in dark blue. In the training cohort, the final minimum cost produced by the policy network gradually decreased exponentially as shown in Figure 3a.

The validation cohort demonstrated similar decreases in cost, reaching a minimum after 4750 iterations. The mean cost in the training buffer monotonically decreased with additional training iterations as shown in Figure 3b, demonstrating that network exploration was continuing to improve the target policy through exploration. Figure 3c shows the policy network loss with training iterations, which decreased exponentially with additional training iterations.

3.2 | Dosimetric comparison with clinical plans

In the 15 test cases, RL plan generation demonstrated mean ± SD execution time of 3.3 ± 0.5 s, which included segment weight optimization. The RL+TPS plan generation required an additional 77.4 ± 5.8 s of automatic iterative optimization per patient within the TPS. Example dose distributions from three representative patients the test cohort are shown in Figure 4. Time required to produce the clinical plans was not controlled, but in a recent comparable cohort of five manually planned localized prostate cancer patients treated with VMAT at our institution, mean ± SD time required for plan optimization was 81.8 ± 56.1 h based on TPS logs, including iterative objective tuning and optimizer execution, similar to the time reported by McIntosh et al. of 115.2 h including 24.7 h for contouring.[34]

All three planning approaches provided conformal treatment plans covering the PTV and preferentially sparing OARs. The RL plans demonstrated some deficiencies with respect to PTV coverage and hot spot, but overall dose distribution characteristics remained similar when refined through the RL+TPS approach, reflecting the fact that the policy network captures the major machine parameter characteristics required to produce a high-quality plan. In all three example cases shown in Figure 4, the RL+TPS approach provided equivalent PTV coverage and comparable PTV D_{max} compared to the clinical plans.

Mean DVH curves for all patients in the test cohort are shown in Figure 5 and associated dose metrics are provided in Table 3. Trends highlighted in Figure 4 are also observed in the mean DVH curves, including relatively lower PTV coverage and higher maximum dose associated with the RL plans, but recovery of PTV coverage and reduction of hot spot and some OAR doses in the RL+TPS plans compared to the clinical plans. The majority of dose metrics demonstrated no statistically-significant difference between the clinical and RL+TPS approaches. The RL+TPS approach provided lower dose with respect to rectum D_{mean}, but

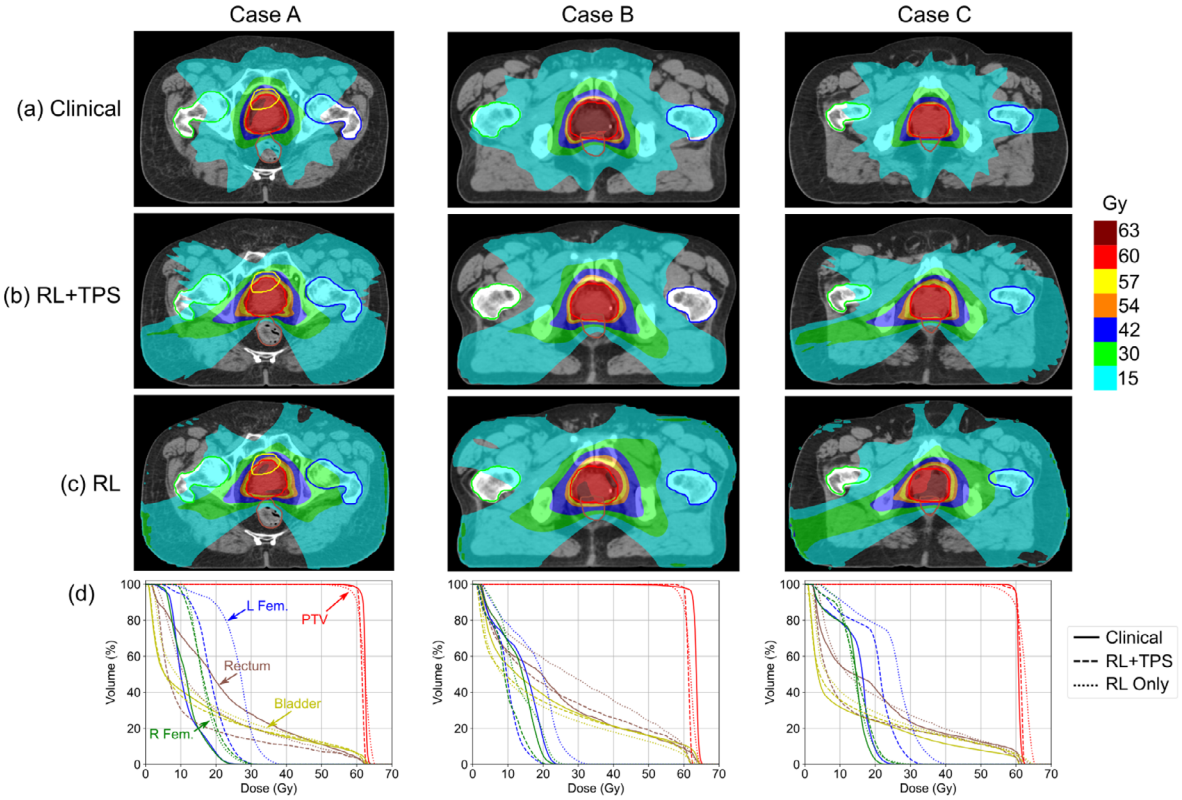

FIGURE 4 (a–c) Example isodose distributions from three example patients in the test cohort produced through (a) clinical planning, (b) the RL+TPS, and (c) the RL approaches. Note that the Clinical and RL+TPS plans were re-normalized to provide the same PTV V60 Gy within each patient. (d) DVH plots corresponding to each case.

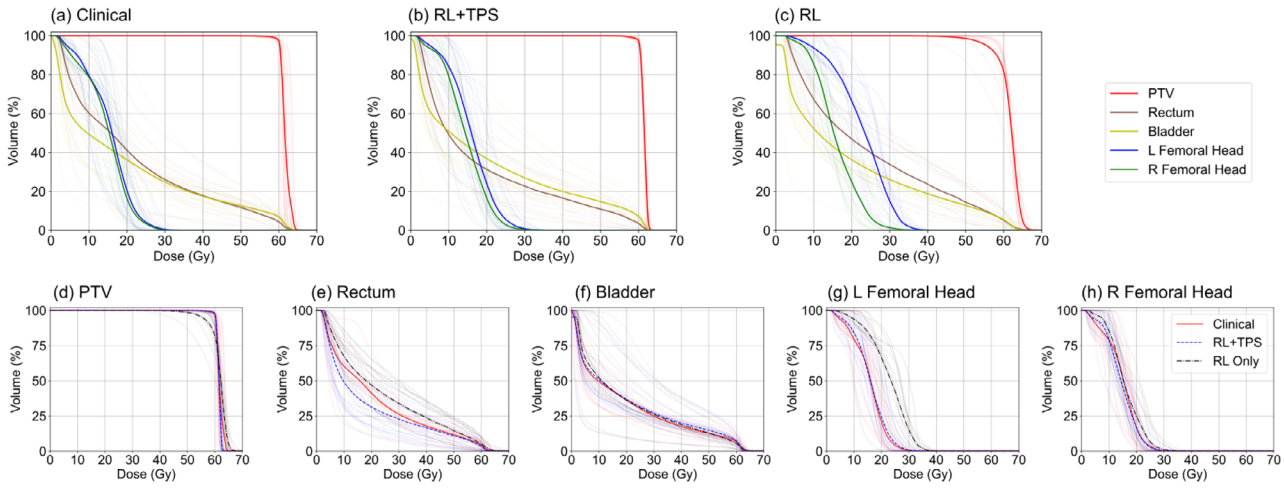

FIGURE 5 Mean DVH curves for patients in the test cohort. (a–c) Mean DVH curves for all structures derived from clinical plans, the RL+TPS approach, and RL approach. (d–h) Mean DVH curves re-plotted on an individual axis for each structure highlighting differences between planning approaches.

slightly higher dose with respect to some points on the bladder DVH. The difference between L and R femoral head doses demonstrates some asymmetry in the RL VMAT plans, potentially attributed to the sequential control point optimization for a single VMAT arc employed in our current RL algorithm.

3.3 | Evaluating policy network produced by supervised learning

Dose metrics associated with the SL and RL VMAT plan sub-analysis are provided in Table 4. Plan quality associated with SL VMAT was significantly poorer compared

TABLE 3 Mean ± SD (p-value) dose metrics for each planning approach in the test cohort.

ROI	Metric	Clinical	RL + TPS	RL
PTV	V6000 (%)	97.2 ± 1.4	97.2 ± 1.4 (p = 1.000)	82.0 ± 11.2 (p = 0.000)*
	D_{min} (cGy)	5701.8 ± 286.1	5730.8 ± 105.1 (p = 0.358)	5056.2 ± 568.4 (p = 0.000)*
	CI	0.833 ± 0.065	0.808 ± 0.031 (p = 0.087)	0.650 ± 0.074 (p = 0.000)*
Body	D_{max} (cGy)	6391.8 ± 153.2	6319.7 ± 55.5 (p = 0.061)	6658.8 ± 118.0 (p = 0.000)*
Rectum	D_{max} (cGy)	6270.3 ± 134.3	6242.9 ± 56.3 (p = 0.218)	6438.4 ± 276.6 (p = 0.018)*
	V6160 (%)	1.7 ± 2.1	1.1 ± 1.1 (p = 0.078)	2.9 ± 2.6 (p = 0.031)*
	V5775 (cc)	4.2 ± 2.5	3.6 ± 2.6 (p = 0.071)	4.6 ± 3.3 (p = 0.208)
	V5390 (%)	9.2 ± 5.1	8.8 ± 5.8 (p = 0.313)	10.7 ± 7.0 (p = 0.065)
	V4620 (%)	14.0 ± 6.8	13.0 ± 7.9 (p = 0.213)	17.3 ± 9.4 (p = 0.017)*
	V3850 (%)	18.9 ± 8.4	17.6 ± 9.6 (p = 0.227)	24.2 ± 11.3 (p = 0.011)*
	V3080 (%)	25.3 ± 10.1	22.2 ± 11.5 (p = 0.089)	32.0 ± 13.0 (p = 0.013)*
	D_{mean} (cGy)	2097.6 ± 604.8	1739.0 ± 741.8 (p = 0.024) *	2367.9 ± 615.1 (p = 0.029)*
Bladder	D_{max} (cGy)	6331.5 ± 143.3	6281.0 ± 51.1 (p = 0.121)	6521.1 ± 95.7 (p = 0.001)*
	V6160(cc)	10.4 ± 8.9	8.6 ± 4.5 (p = 0.252)	10.5 ± 6.8 (p = 0.475)
	V5775 (%)	8.8 ± 3.5	9.9 ± 4.4 (p = 0.016)*	7.3 ± 4.4 (p = 0.046)*
	V5390 (%)	10.9 ± 4.3	12.7 ± 5.5 (p = 0.004)*	10.1 ± 5.3 (p = 0.208)
	V4620 (%)	14.3 ± 5.7	16.8 ± 7.2 (p = 0.002)*	14.7 ± 6.8 (p = 0.362)
	V3850 (%)	18.5 ± 7.5	20.9 ± 8.9 (p = 0.007)*	19.4 ± 8.5 (p = 0.214)
	V3465 (%)	21.0 ± 8.8	23.4 ± 10.1 (p = 0.016)*	22.1 ± 9.5 (p = 0.220)
L femoral head	D_{max} (cGy)	2717.7 ± 322.9	2786.7 ± 417.6 (p = 0.291)	3609.7 ± 352.0 (p = 0.000)*
	D40% (cGy)	1688.2 ± 304.5	1726.1 ± 311.6 (p = 0.376)	2516.4 ± 365.9 (p = 0.000)*
R femoral head	D_{max} (cGy)	2608.6 ± 340.8	2695.6 ± 409.8 (p = 0.226)	2916.0 ± 428.3 (p = 0.008)*
	D40% (cGy)	1625.6 ± 303.1	1598.0 ± 256.3 (p = 0.400)	1727.9 ± 308.7 (p = 0.188)

Metrics with statistically significant differences from the clinical plans are indicated by asterisks.

TABLE 4 Mean ± SD dose metrics for the SL and RL VMAT approaches in the test cohort.

Plan Approach	PTV V6000 (%)	CI	Body D_{max} (cGy)	Rectum D_{mean} (cGy)
RL VMAT	82.0 ± 11.2	0.650 ± 0.074	6658.8 ± 118.0	2367.9 ± 615.1
SL VMAT	59.5 ± 17.7	0.485 ± 0.158	6918.9 ± 106.8	1884.1 ± 562.5

All metric differences between RL and SL were statistically significant ($p < 0.001$).

to the RL VMAT plans, where PTV coverage and CI were significantly lower and hot spot was significantly higher. Mean rectum dose provided by SL VMAT was lower at the expense of significantly reduced PTV coverage. This comparison demonstrates that final policy network performance is driven by the RL component rather than SL.

4 | DISCUSSION

We have demonstrated an approach for VMAT MPO using RL that generates machine parameters for a clinical linac beam model in under 4 s, and a workflow that combines the RL VMAT approach with a commercial TPS providing high-quality and deliverable VMAT plans for localized prostate cancer that may be reviewed and optimized further if needed. The combined RL+TPS automatically created high-quality and deliverable VMAT plans for all 15 patients in the independent test cohort, and provided lower OAR dose for equivalent target coverage in many cases. In contrast to most automatic planning approaches, the RL VMAT approach replaces dose objective prediction and numerical VMAT MPO with a single algorithmic agent which directly predicts machine parameters based on an input planning CT and contours. Through the policy gradient approach, cost function definition and tuning are completely handled during RL training and incorporated in the policy network weights. Specifically, the policy network is trained to minimize a DVH-based cost function for each patient as evaluated during training, but no longer requires explicit cost function evaluation once trained. In this preliminary study, we employed a simple default cost function during RL training without patient-specific

parameter tuning. However, the policy gradient RL approach creates the potential to incorporate arbitrary cost functions during training, including patient-specific cost function tuning through RL with human feedback (RLHF),[21] similar to the approach used to fine tune GPT-4.[20] The RL+TPS approach currently requires manual plan refinement using conventional iterative optimization in instances where physicians request plan changes. This feedback could be used to drive an RLHF approach to further fine-tune the policy network, leading to better performing or physician-specific policy networks. Moreover, our RL VMAT MPO approach could be combined with previously proposed RL-based dose objective tuning approaches such as the VTPN[24] to provide further performance improvements.

Supervised learning (SL) was used to initialize network weights for RL, similar to approaches applied to other complex problems with high dimensional action spaces and sparse rewards.[18] We compared the quality of VMAT plans produced using the policy network weights resulting from the SL-based initialization alone to those produced through RL, and found significantly improved plan quality achieved following the full RL training. We associate this improvement in performance to multiple aspects of RL training including improved plan quality in the transition buffer achieved through exploration and an increased number of training iterations, among other potential factors.

While the RL+TPS plans demonstrated comparable dosimetry to clinical plans, the RL plans demonstrated some major dosimetric limitations with respect to PTV dose which would prevent clinical deployment without refinement. Specifically, target coverage was too low and overall plan hot spot was too high for cases in the test set. There are several contributors to these performance characteristics which could be overcome through further investigation. The first limitation is the dependence on a default cost function during RL training, which did not incorporate patient-specific objective tuning as is required in conventional inverse VMAT optimization, so may have encouraged these dosimetric characteristics during training. Another limitation is our current training cohort size and simple data augmentation approach, which has not fully overcome challenges in network over-fitting to the training set which limits performance on new prospective cases, including the test cohort. Over-fitting is apparent in Figure 3a,b, where the cost associated with the validation cohort remains higher than the test cohort for the latter portion of training iterations, although continues to decrease. High RL performance requires a very large number of iterations to effectively explore the parameter space, particularly for high-dimensional problems; however, this large number of training iterations can lead to over-fitting to patients in the training set. This type of over-fitting is a smaller issue in other RL applications such as board games where the board is identical during training and deployment. While we took several steps to mitigate over-fitting in this study including increasing training cohort size, implementing data augmentation, and limiting policy network complexity, policy network over-fitting remains a central challenge to the RL approach. Since RL training does not require existing VMAT plans for all training cases, there are many potential strategies to significantly increase training data size that we are investigating.

This study has several limitations. In our comparison of RL VMAT with clinical planning, we employed simple comparison of dose metrics which were averaged across patients. As previously mentioned, we did not incorporate extensive cost function tuning for either RL training or the RL+TPS approach, with values provided in Tables 1 and 2, which contributed to the dosimetric trade-offs observed in our test cases. For example, the lower overall maximum dose observed in the RL+TPS plans could be relaxed by modifying the cost function terms, potentially enabling further reductions in OAR dose. Previous studies have shown that clinician preference for treatment plans is not always fully characterized by DVH objectives, and requires full review of the plan details in the context of treatment intent,[34] so we cannot conclude the RL+TPS plans would be clinically acceptable without further detailed review or blinded reader study, which was beyond the scope of the present report but is planned as a follow-up study. The findings of detailed prospective review may suggest areas for improvement, which again could be implemented through cost function modifications during RL training. Another limitation of the present study is our exclusive focus on VMAT for localized prostate cancer, which allowed for reduction in the number of active MLC leaves due to target size limits, and further limited the anatomical and plan complexity observed in the training and test cohorts. This treatment site was deliberately chosen to enable development of the RL VMAT approach and represents a benchmark for automated planning. The RL VMAT approach could be extended to more complex treatment sites through network re-training and other potential algorithmic refinements and will be the topic of future investigations. For example, this study focused on single-phase single arc plans with fixed collimator angles. Multi-phase and multi-arc planning is possible with the RL approach, but will require modification of the policy network, state arrays, and network re-training. Finally, the RL VMAT algorithm depends on many hyperparameters that could be optimized further including the number of seed plans used to initialize training, the policy network design and all network training parameters, and are the topic of ongoing investigations.

5 | CONCLUSIONS

We have developed an approach for VMAT MPO using RL, which was applied to the optimization of VMAT

plans for localized prostate cancer for a clinical linac. We also demonstrated an approach that combines RL with a clinical TPS, enabling plan review and refinement as is required for potential clinical deployment. The RL+TPS approach automatically produced deliverable VMAT plans with comparable dosimetry to clinical plans. The RL VMAT approach shows promise to achieve fast and high-quality auto-planning, potentially exceeding the performance of existing algorithms. Limitations of the current algorithm were identified including network over-fitting, and directions for future investigation were identified and discussed.

ACKNOWLEDGMENTS

This project was funded by the Commonwealth Fund.

CONFLICT OF INTEREST STATEMENT

William T. Hrinivich, Xun Jia, and Junghoon Lee conduct research funded by Varian Medical Systems unrelated to this study. The remaining authors have no relevant conflicts of interest to disclose.

REFERENCES

1. Palma D, Vollans E, James K, et al. Volumetric modulated arc therapy for delivery of prostate radiotherapy: comparison with intensity-modulated radiotherapy and three-dimensional conformal radiotherapy. *Int J Radiat Oncol Biol Phys*. 2008;72(4):996-1001. doi:10.1016/j.ijrobp.2008.02.047
2. Unkelbach J, Bortfeld T, Craft D, et al. Optimization approaches to volumetric modulated arc therapy planning. *Med Phys*. 2015;42(3):1367-1377. doi:10.1118/1.4908224
3. MacFarlane M, Hoover DA, Wong E, Battista JJ, Chen JZ. Technical note: a fast inverse direct aperture optimization algorithm for volumetric-modulated arc therapy. *Med Phys*. 2020;47(4):1558-1565. doi:10.1002/mp.14074
4. Men C, Romeijn HE, Jia X, Jiang SB. Ultrafast treatment plan optimization for volumetric modulated arc therapy (VMAT). *Med Phys*. 2010;37(11):5787-5791. doi:10.1118/1.3491675
5. Peng F, Jia X, Gu X, Epelman MA, Romeijn HE, Jiang SB. A new column-generation-based algorithm for VMAT treatment plan optimization. *Phys Med Biol*. 2012;57(14):4569-4588. doi:10.1088/0031-9155/57/14/4569
6. Wang C, Zhu X, Hong JC, Zheng D. Artificial intelligence in radiotherapy treatment planning: present and future. *Technol Cancer Res Treat*. 2019;18:1533033819873922. doi:10.1177/1533033819873922
7. Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: a review of data-driven approaches. *Med Phys*. 2019;46(6):2760-2775. doi:10.1002/mp.13526
8. Momin S, Fu Y, Lei Y, et al. Knowledge-based radiation treatment planning: a data-driven method survey. *J Appl Clin Med Phys*. 2021;22(8):16-44. doi:10.1002/acm2.13337
9. Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys*. 47(2):297-306. doi:10.1002/mp.13896. Published online December 21, 2018.
10. Murakami Y, Magome T, Matsumoto K, Sato T, Yoshioka Y, Oguchi M. Fully automated dose prediction using generative adversarial networks in prostate cancer patients. *PLoS One*. 2020;15(5):e0232697. doi:10.1371/journal.pone.0232697
11. Nilsson V, Gruselius H, Zhang T, DeKerf G, Claessens M. Probabilistic dose prediction using mixture density networks for automated radiation therapy treatment planning. *Phys Med Biol*. 2021;66(5):055003. doi:10.1088/1361-6560/abdd8a
12. Nguyen D, McBeth R, Sadeghnejad Barkousaraie A, et al. Incorporating human and learned domain knowledge into training deep neural networks: a differentiable dose-volume histogram and adversarial inspired framework for generating Pareto optimal dose distributions in radiation therapy. *Med Phys*. 2020;47(3):837-849. doi:10.1002/mp.13955
13. Li X, Zhang J, Sheng Y, et al. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning. *Phys Med Biol*. 2020;46(6):E256-E256. doi:10.1088/1361-6560/aba5eb
14. Vandewinckele L, Willems S, Lambrecht M, Berkovic P, Maes F, Crijns W. Treatment plan prediction for lung IMRT using deep learning based fluence map generation. *Physica Medica*. 2022;99:44-54. doi:10.1016/j.ejmp.2022.05.008
15. Wang W, Sheng Y, Palta M, et al. Deep learning–based fluence map prediction for pancreas stereotactic body radiation therapy with simultaneous integrated boost. *Adv Radiat Oncol*. 2021;6(4):100672. doi:10.1016/j.adro.2021.100672
16. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT press; 2018.
17. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550(7676):354-359. doi:10.1038/nature24270
18. Vinyals O, Babuschkin I, Czarnecki WM, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. 2019;575(7782):350-354. doi:10.1038/s41586-019-1724-z
19. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518(7540):529-533. doi:10.1038/nature14236
20. OpenAI. GPT-4 Technical Report. Published online March 15, 2023. http://arxiv.org/abs/2303.08774
21. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Published online June 12, 2017. http://arxiv.org/abs/1706.03741
22. Shen C, Gonzalez Y, Klages P, et al. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Phys Med Biol*. 2019;64(11):115013. doi:10.1088/1361-6560/ab18bf
23. Shen C, Chen L, Jia X. A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. *Phys Med Biol*. 2021;66(13):134002. doi:10.1088/1361-6560/ac09a2
24. Shen C, Nguyen D, Chen L, et al. Operating a treatment planning system using a deep-reinforcement-learning based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. *Med Phys*. 2020;47(6):2329-2336. doi:10.1002/mp.14114. Published online March 5, 2020:mp.14114.
25. Hrinivich WT, Lee J. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. *Med Phys*. 2020;47(12):6140-6150. doi:10.1002/mp.14544
26. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In *International conference on machinelearning*, pp. 387-395. Pmlr, 2014.
27. Bedford JL, Thomas MDR, Smyth G. Beam modeling and VMAT performance with the agility 160-leaf multileaf collimator. *J Appl Clin Med Phys*. 2013;14(2):172-185. doi:10.1120/jacmp.v14i2.4136
28. Jacques R, Wong J, Taylor R, McNutt T. Real-time dose computation: gPU-accelerated source modeling and superposition/convolution. *Med Phys*. 2011;38(1):294-305. doi:10.1118/1.3483785
29. Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Med Phys*. 1989;16(4):577-592.

30. Dearnaley D, Syndikus I, Mossop H, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. *Lancet Oncol*. 2016;17(8):1047-1060. doi:10.1016/S1470-2045(16)30102-4
31. Kingma DP, Ba JL. Adam: a method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; 2015. http://arxiv.org/abs/1412.6980
32. Lillicrap TP, Hunt JJ, Pritzel A, et al. Continuous control with deep reinforcement learning. Published online September 9, 2015. http://arxiv.org/abs/1509.02971
33. Paddick I. A simple scoring ratio to index the conformity of radiosurgical treatment plans. Technical note. *J Neurosurg*. 2000;93(3):219-222. doi:10.3171/jns.2000.93.supplement
34. McIntosh C, Conroy L, Tjong MC, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med*. 2021;27(6):999-1005. doi:10.1038/s41591-021-01359-w

How to cite this article: Hrinivich WT, Bhattacharya M, Mekki L, et al. Clinical VMAT machine parameter optimization for localized prostate cancer using deep reinforcement learning. *Med Phys*. 2024;51:3972–3984. https://doi.org/10.1002/mp.17100

APPENDIX A: RL DOSE ENGINE BEAM MODELING

A beam model was defined for the in-house GPU-based collapsed cone convolution (CCC) super-position photon dose engine used for RL training, matching the 6 MV beam of a Versa HD linac (Elekta, Stockholm, Sweden) at our institution. Beam model parameters, including x-ray spectrum, beam profiles, off-axis softening, and electron contamination, were manually tuned to match depth dose curves, output factors, lateral X-profiles, and lateral Y-profiles of the commissioning data in our clinical TPS. Output factors for various square field sizes (1, 2, 3, 5, 10, 15, 20, 30, and 40 cm) were matched at 10 cm depth. Example depth dose curves and lateral profiles comparing the RL dose engine with our clinical TPS (RayStation 2023B, RaySearch, Stockholm, Sweden) in a virtual water phantom for 100 MU and 90 cm source to surface distance (SSD) are provided in Figure A1a,b. An example VMAT plan for a prostate cancer patient planned in the clinical TPS and re-computed using the RL dose engine is shown in Figure A1c,d. The gamma pass rate between the RL dose engine and TPS for this example patient was 96.6% with 3%/2 mm criteria. While the dose engines show some residual differences attributed to limitations in the simplified RL dose engine model, the accuracy was found sufficient for RL training and use in the RL+TPS approach, where the final dose is computed using the clinical dose engine for final plan approval.

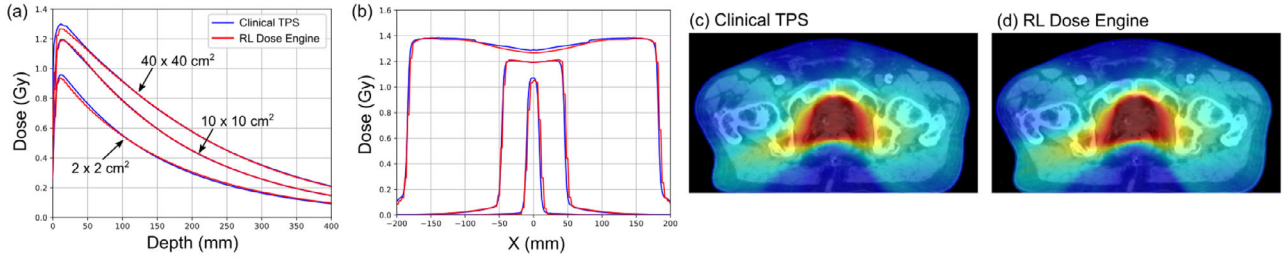

FIGURE A1 Example depth dose curves (a) and lateral profiles at 1.5 cm depth (b) for 2×2 cm^2, 10×10 cm^2, and 40×40 cm^2 square fields at 90 cm SSD computed in a virtual water phantom using the commercial TPS and the RL dose engine for 100 MU. Example VMAT prostate plan produced in the commercial TPS (c) and re-computed using the RL dose engine (d).