Received: 3 May 2024 | Revised: 18 July 2024 | Accepted: 1 August 2024

DOI: 10.1111/aor.14845

Artificial Organs WILEY

MAIN TEXT

Physiological control for left ventricular assist devices based on deep reinforcement learning

Diego Fernández-Zapico | Thijs Peirelinck[1] | Geert Deconinck[1] | Dirk W. Donker[2,3] | Libera Fresiello[2]

[1]Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

[2]Cardiovascular and Respiratory Physiology, University of Twente, Enschede, the Netherlands

[3]Intensive Care Center, University Medical Center Utrecht, Utrecht, the Netherlands

Correspondence
Diego Fernández-Zapico.
Email: d.fernandez.zapico@gmail.com

Abstract

Background: The improvement of controllers of left ventricular assist device (LVAD) technology supporting heart failure (HF) patients has enormous impact, given the high prevalence and mortality of HF in the population. The use of reinforcement learning for control applications in LVAD remains minimally explored. This work introduces a preload-based deep reinforcement learning control for LVAD based on the proximal policy optimization algorithm.

Methods: The deep reinforcement learning control is built upon data derived from a deterministic high-fidelity cardiorespiratory simulator exposed to variations of total blood volume, heart rate, systemic vascular resistance, pulmonary vascular resistance, right ventricular end-systolic elastance, and left ventricular end-systolic elastance, to replicate realistic inter- and intra-patient variability of patients with a severe HF supported by LVAD. The deep reinforcement learning control obtained in this work is trained to avoid ventricular suction and allow aortic valve opening by using left ventricular pressure signals: end-diastolic pressure, maximum pressure in the left ventricle (LV), and maximum pressure in the aorta.

Results: The results show controller obtained in this work, compared to the constant speed LVAD alternative, assures a more stable end-diastolic volume (*EDV*), with a standard deviation of 5 mL and 9 mL, respectively, and a higher degree of aortic flow, with an average flow of 1.1 L/min and 0.9 L/min, respectively.

Conclusion: This work implements a deep reinforcement learning controller in a high-fidelity cardiorespiratory simulator, resulting in increases of flow through the aortic valve and increases of *EDV* stability, when compared to a constant speed LVAD strategy.

KEYWORDS

cardiorespiratory simulator, deep reinforcement learning, heart failure, left ventricular assist device, physiological control

1 | INTRODUCTION

Heart failure (HF) is one of the leading causes of morbidity and mortality worldwide[1] and can arise due to different underlying cardiovascular and noncardiovascular diseases.[2] Although pharmacological therapy is a mainstay of management in HF patients, the disease can progress to a level of severity that only heart transplantation would remain a therapeutic option. However, due to a tremendous shortage of donor organs, only a limited number of patients can benefit from transplantation. Most patients with end-stage HF are therefore supported with left ventricular assist devices (LVADs) be it as a bridge to transplantation or destination therapy. These long-term LVADs represent a technology based on implantable rotary blood pumps that support the failing left ventricle (LV). LVADs drain blood from the LV and pump it into the ascending aorta and systemic circulation. From a technical point of view, LVADs work in parallel with the native ventricle, creating an artificial blood circuit in addition to the native one through the aortic valve. Recent improvements and miniaturization of LVAD technology[3] also brought the need to tailor the level of support flow to the patient's specific needs ideally based on a dedicated controller that automatically adjusts the LVAD rotational speed.

An improved control of the LVAD speed is expected to elicit clinical benefits, compared to the current situation with the LVADs operating at a constant speed requiring manual adjustments. Depending on how the LVAD speed control is designed, it can prevent adverse events, such as LV suction or collapse and prolonged aortic valve closure. All in all, an automatic physiological LVAD control has the potential to create better hemodynamic conditions that could reduce pump thrombosis and cerebrovascular events.[4] The design and testing of an LVAD control is challenging and involves many aspects related to its safety and effectiveness. The LVAD automatic control should be able to adapt the device flow according to the level of filling of the LV (preload), the level of perfusion needed by the organs (target cardiac output), and additionally desired conditions, for example, opening of the aortic valve, enhancing of systemic pressure pulsatility. Some of these conditions might translate into technical specifications in contradiction with each other; therefore, a control might not satisfy all of them but only those considered the most relevant for the target population. At present, several attempts have been conducted, with LVAD controllers based on Proportional Integral Derivative (PID) and Fuzzy Logic, among others.[5] However, only a few controllers have been successfully tested in in vitro[6] or in vivo trials.[7,8]

A promising, emerging technique for the development of control systems in medical technology is Reinforcement Learning (RL), a family of Machine Learning (ML) based on data driven control techniques. Previous work by Li et al.,[9] has shown that LVAD controllers based on RL can be quicker and more effective than traditional PID controllers. However, LVAD controllers based on RL have not extensively been researched, so with this work, we provide an example of applicability of this technique.

1.1 | RL in healthcare

RL learns a way of taking actions in a given environment by interacting with it. Coronato et al.,[10] studied the impact of RL in a wide range of applications in the healthcare domain. The most closely related RL application in control systems involves the so-called "training an agent (the controller) in an environment (the medical device to be controlled)." The agent must learn to act by training in a simulated environment, before being tested in vivo. Thus, it is critical that the controller is trained in a realistic and flexible, simulated environment, ideally validated with real life patient data.

An example of RL control is the artificial pancreas (AP) that tunes insulin infusion to maintain blood glucose levels within physiological ranges. RL controllers for the AP can assure a better performance and a higher level of personalization than traditional PID and model predictive control techniques, when trained on a reliable set of patient data or in a high-fidelity FDA-approved simulator.[11–13] However, RL controllers have two major drawbacks: the controller is a black box and its training is time inefficient. The latter can be improved with parallel training approaches[10] or, alternatively, simplifying the problem formulation: discretizing actions, avoiding sparse rewards, and shaping the state space according to the reward.[14]

Encouraging results have also been achieved in RL controls for mechanical ventilation and sedation dosage in intensive care units, either using historical patient data or a model of the dynamics of the patient for training.[15,16]

In the LVAD domain, most effort has gone into complementing existing controllers with supervised learning, a subgroup of ML that aims to learn patterns from labeled data in order to make predictions in unlabeled sets. Fetanat et al.,[17] combined a convolutional neural network and model-free adaptive control to achieve a sensorless preload-based controller using the LVAD flow as the only input. The controller is able to perform similarly to a traditional sensor-based PID controller. In the context of biventricular assisted devices (BiVAD), Ng et al.,[18] combined a multiobjective neural network based predictive controller and a Frank–Starling controller for a BiVAD. The controller achieves robust performance without hemodynamic instabilities, unlike the dual-independent Frank–Starling controller previously used.

Recently, a controller that uses a soft-actor-critic RL agent,[9] was developed for LVAD technology. The RL agent was trained by interacting with a physiological model of the cardiovascular system supported by an LVAD. Although a relatively simple cardiovascular simulator was used, the approach presented resulted in a controller that can respond more quickly and effectively than the traditional PID controllers. The approach of using a cardiovascular simulator allows to reproduce an enormous variety of conditions and patient profiles that can be used both for training and testing of the RL controller. In this article, we explore and further develop this methodology and analyze the final RL controller outcome.

Our work follows the approach of designing a physiological LVAD controller, based on the RL techniques and using the data of a high-fidelity cardiorespiratory simulator simulating different states of a HF patient under LVAD support. A physiological LVAD controller has the general aim to deliver a flow according to patient's circulatory needs. Although this medical objective might appear straightforward, it constitutes a complex matter as it involves an intricate series of relationships among different components of the cardiovascular system. In this work, we focused the controller on two main aspects: LV suction prevention and aortic valve opening. LV suction events are common complications occurring in LVADs when the pump speed is too high for the amount of LV preload, and thus, the LVAD drains too much blood and the LV chamber tends to collapse. LV suction can cause chest pain and cardiac tissue damage[4] as well as low flow through the LVAD.[4] The functionality of the aortic valve is another point of concern during LVAD therapy, as a too high LVAD speed might reduce LV preload to a point that the LV ventricle does not develop a systolic pressure able to open the aortic valve for multiple, consecutive heart beats. Prolonged aortic valve closure has been linked to regurgitation rendering the LVAD less effective[4] and notably increased risk of thrombosis.[19]

2 | METHODS

2.1 | Deep reinforcement learning

2.1.1 | Finite MDPs

It is common to formulate discrete-time stochastic control problems as Markov Decision Processes (MDP). In a finite MDP, at each time step, an agent takes actions $A_t \in \mathcal{A}$ in a dynamic environment and, in return, has access to its states or observations $S_t \in \mathcal{S}$ and rewards $R_t \in \mathscr{R}$, where all sets have a finite number of elements.[20]

The dynamics or transitions of the environment at any step t are only conditioned on the previous action and state, following the Markov property:

$$p(s', r \mid s, a) = P\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}$$

with s and s' being states at step $t-1$ and t, respectively; a actions taken at step $t-1$; and r the reward associated with the transition.

The agent aims to maximize present and future rewards r, giving more importance to early rewards by discount factor γ. This can be achieved with a cumulative reward, the expected discounted return, which can be written recursively:

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k = R_{t+1} + \gamma G_{t+1}$$

where $0 \leq \gamma \leq 1$ and T is the time of termination of the episode.

A stochastic policy π is a model followed by the agent to take actions in a given state: $a \sim \pi(\cdot \mid s)$. Finding policies that maximize G_t relies on methods that can accurately estimate the value function V^π and the action-value function Q^π:

$$V^\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right]$$

$$Q^\pi(s, a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

where $\mathbb{E}_\pi$ is the expectation under policy π.

2.1.2 | Proximal policy optimization

Policy gradient methods aim to directly learn a policy π_θ from the environment. This approach avoids dealing with the intermediate step of learning action-value functions. These methods use a loss function defined as follows:

$$L^{PG}(\theta) = \mathbb{E}_t\left[\log \pi_\theta\left(a_t \mid s_t\right) \widehat{A}_t\right]$$

where $\widehat{A}_t$ is the estimator of the advantage function $A^\pi(s,a) = Q^\pi(a,s) - V^\pi(s)$.[21] It controls the direction of change given by the gradient $g = \mathbb{E}_t\left[\nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right) \widehat{A}_t\right]$.

The Proximal Policy Optimization (PPO) algorithm is a model-free on-policy approach that belongs to the family of trust region methods. These methods propose to maximize a surrogate of the loss function, constraining large policy updates to achieve robust learning performance.[21] In particular, PPO[22] proposes a clipped surrogate objective function, defined as follows:

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(\left(r_t(\theta)\right)\widehat{A}_t, clip\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\widehat{A}_t\right)\right]$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is a probability ratio between the current and the past policy and ϵ is the clip range controlling the changes. This loss function takes the minimum of the regular and the clipped component, controlling the range of values allowed for L^{CLIP} and, by extension, of its gradient.

In the Stable Baselines3 implementation, the loss function includes one more term related to the estimation of $V_\theta(s)$. It is worth noting that $V_\theta(s)$ is estimated using a neural network that shares parameters with the policy network π_θ. The complete loss function used in this project is defined as follows:

$$L^{CLIP+VF} = \mathbb{E}\left[L^{CLIP}(\theta) - c_{\text{val}}L^{VF}(\theta)\right]$$

where $L^{VF}(\theta) = \left(V_\theta(s) - V^{\text{target}}\right)^2$ and c_{val} is a constant that weighs the L^{VF} component.

This algorithm is based on batch reinforcement learning.[23] Using the current policy, the agent collects transitions of length T, known as the horizon, which is recommended to be smaller than the episode length. After T timesteps, the estimates $\widehat{A}_t$ are computed to perform gradient updates on $L^{CLIP+VF}$ using the Adam optimizer. The Stable Baselines3 library is used for the implementation of the PPO agent,[24] with most hyperparameters set to their defaults except for T, ϵ, the batch size (b), and the learning rate (lr), which are adjusted to optimize learning.

2.2 | Cardiorespiratory system-VAD simulator

A high-fidelity cardiorespiratory system developed in LabVIEW was used for the work. The simulator includes a lumped parameter model of the closed-loop circulation, ventilation mechanics, gas exchange in the lungs and peripheral tissues, baroreflex control, metabolic peripheral control, and ventilation control. The simulator was validated when representing healthy and HF patients with LVAD therapy against clinical data.[25,26] Moreover, the simulator can be tuned to reproduce different patient conditions by adjusting specific cardiovascular and respiratory parameters.[27]

For the purpose of this article, only the cardiovascular part of the simulator was used, and the model was tuned to reproduce a generic HF patient. Starting from this baseline condition, cardiovascular parameters were changed sequentially to capture a large hemodynamic variability, to mimic a broad spectrum of patient conditions. Namely, the parameters changed were as follows: total blood volume (*TBV*), heart rate (*HR*), systemic vascular resistance (*SVR*), pulmonary vascular resistance (*PVR*), and left ventricular end-systolic elastance (*Els*), and right ventricular end-systolic elastance (*Ers*).

The simulator is a deterministic model; therefore, given a combination of cardiovascular parameters and LVAD settings, the output produced in terms of pressures, flows, and volumes is consistently the same. The LVAD is modeled as a third generation centrifugal pump, similar to the HeartWare VAD (HVAD), with a range of pump speeds between 1800 and 4020 rpm.

2.3 | Deep reinforcement learning controller

In this work, the state-action-reward design aimed to avoid two main challenges in VAD operation: LV suction and aortic valve closure. The control takes as main input the LV preload, defined here as the diastolic LV pressure (*EDP*). This MDP is formulated in a beat-to-beat manner with t representing the step or heart cycle. Additionally, this is done following the guidelines of Gym environments,[28] with a connection layer between LabVIEW and Python (see Figure 1).

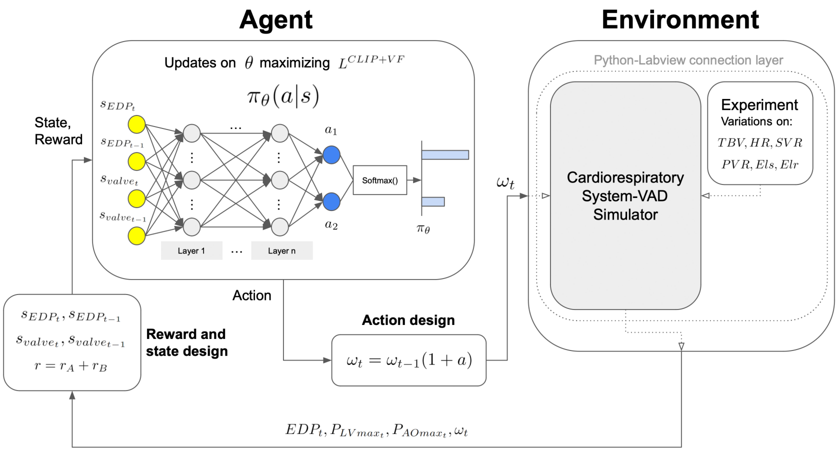

FIGURE 1 Agent–environment interaction for the proposed RL controller based on PPO. [Color figure can be viewed at wileyonlinelibrary.com]

The first and main part of the reward function r_A aims to follow the Frank–Starling mechanism. The focus of this implementation is on changing the speed depending on the value of *EDP*:

$$r_A = \begin{cases} 1, & s_{EDP_t} a \geq 0 \\ -2, & s_{EDP_t} a < 0 \end{cases}$$

where the observation $s_{EDP_t} = EDP_t - EDP_{\text{ref}}$ and the constant EDP_{ref} can be set by physicians and is part of the hyperparameter tuning of this project. Actions a are discretized as follows:

$$a = \begin{cases} 0.01 \\ -0.01 \end{cases}$$

$$\omega_t = \omega_{t-1}(1 + a)$$

The second part of the reward function r_B concerns avoiding aortic valve closure. An observation related to this result can be defined as the pressure difference in systole:

$$s_{\text{valve}_t} = P_{LV\text{max}_t} - P_{AO\text{max}_t}$$

where $P_{LV\text{max}}$ is the maximum pressure in the LV and $P_{AO\text{max}}$, the maximum pressure in the aorta. When s_{valve} is positive, valve opening is taking place, and the agent is rewarded when the speed is increased. By the nature of the problem, the agent is not encouraged to infinitely increase the speed, as both s_{EDP} and s_{valve} are reduced with speed increases. In case of $s_{\text{valve}} < 0$ (valve closure), just like in the case of controlling the *EDP*, the agent should decrease the speed to achieve the desired overlap of pressure signals:

$$r_B = \begin{cases} 1, & s_{\text{valve}_t} a \geq 0 \\ 0, & s_{\text{valve}_t} a < 0 \end{cases}$$

The final reward function is then defined as follows:

$$r = r_A + r_B$$

where the first term (r_A) has preference over the second one (r_B), as the reward given for not following the preferred policy is more negative in the former.

It is worth noting that, by default, the Markov Property of this MDP does not hold due to the influence of nonobservable experiment parameters. As an attempt to restore the Markov Property, lagged states were included to the state vector:

$$s = \left(s_{EDP_t}, s_{EDP_{t-1}}, s_{\text{valve}_t}, s_{\text{valve}_{t-1}} \right)$$

TABLE 1 Recorded parameters[a] for the baseline HF patient with a LVAD running in a range of 2300–2500 rpm.

Parameter	Units	Value
HR	bpm	80
TCO	L/min	4.9–5.1
Q_{LV}	L/min	0.7–1.2
WP	mmHg	14–16
CVP	mmHg	7
PAP	mmHg	22–23
PAPES	mmHg	27–28
PAPED	mmHg	17–19
AOP	mmHg	82–86
AOPES	mmHg	98–100
AOPED	mmHg	68–75
LVES	mL	129–131
LVED	mL	179–183

[a] Heart rate (HR), total cardiac output (TCO), mean LV flow $\left(Q_{LV}\right)$, mean wedge pressure (WP), mean central venous pressure (CVP), mean pulmonary arterial pressure (PAP), end-systolic PAP $(PAPES)$, end-diastolic PAP $(PAPED)$, mean aortic pressure (AOP), end-systolic AOP $(AOPES)$, end-diastolic AOP $(AOPED)$, end-systolic LV volume $(LVES)$, and end-diastolic LV volume $(LVED)$.

2.4 | Training and evaluation

The simulation procedure consists of setting the simulator to a baseline HF patient with an LVAD running in a range of 2300–2500 rpm, matching the hemodynamic conditions of a patient at rest in a study by Fresiello et al.[29] The baseline value of *Els* was chosen to reproduce a residual contractility, so to assure aortic valve opening at baseline condition (see Table 1).

During the evaluation episode, six cardiovascular parameters are modified sequentially: starting from the baseline value each parameter is changed within a minimum and a maximum value while keeping the rest of the parameters set to their baseline values, as reported in Table 2. Each change of parameter is followed by a stabilization phase (see *Upper limit* curve in Figure 2). A time window of 10 s is used to increase and decrease *HR* while 50 s for the other parameters, according to the time constants reported in a study by Ursino et al.[30] The length of the evaluation episode is 900 s or 1206 heart cycles.

In the training phase, the parameter order and the range of its change are randomized. The complete training process is composed of training episodes with a length of 150 s or 201 heart cycles. Each training episode contains changes on individual parameters chosen at random and for an amount of change randomly chosen between an upper and lower limit (see maximum and minimum range in Table 2 and *Upper*

TABLE 2 Maximum and minimum ranges of change and default values for parameters.[a]

Parameter	Units	Default	Maximum range	Minimum range
TBV	mL	5800	5300–5800	5425–5800
HR	bpm	80	60–180	60–150
SVR	mmHgs/cm^3	0.9	0.4–1.4	0.525–1.275
PVR	mmHgs/cm^3	0.054	0.054–0.1	0.054–0.075
Ers	mmHg/cm^3	0.6	0.2–1	0.3–0.9
Els	mmHg/cm^3	0.8	0.4–1.2	0.5–1.1

[a] Total blood volume (TBV), heart rate (HR), systemic vascular resistance (SVR), pulmonary vascular resistance (PVR), right ventricular end-systolic elastance (Ers), and left ventricular end-systolic elastance (Els).

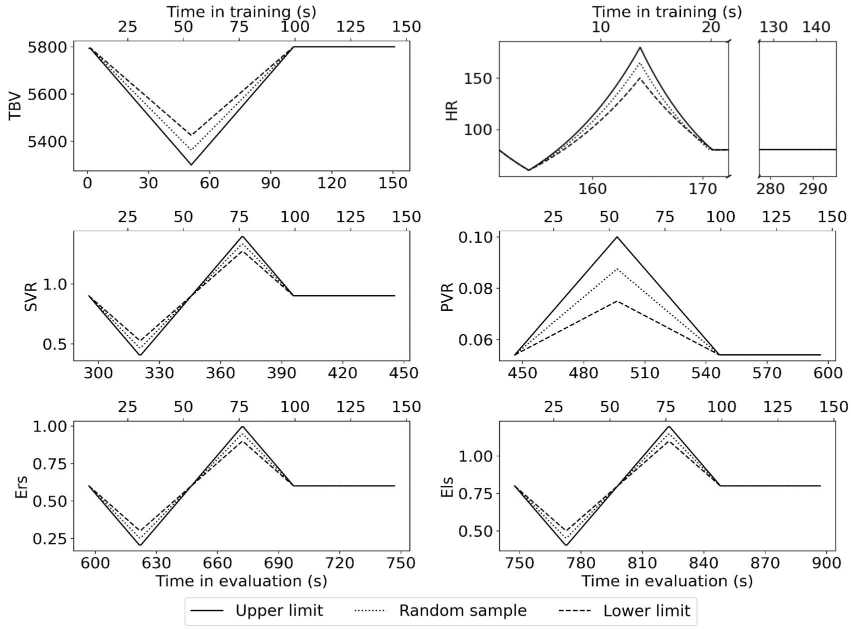

FIGURE 2 Summary of experiment conditions during training and evaluation. In order from top left to bottom right: Change of total blood volume (TVB), change of heart rate (HR), change of systemic vascular resistance (SVR), change of pulmonary vascular resistance (PVR), change of right ventricular end-systolic elastance (Ers), and change of left ventricular end-systolic elastance (Els).

limit and *Lower limit* lines in Figure 2). The concatenation of all training episodes results in a random order of change of experiment parameters during the training process.

The training strategy is composed of two steps: firstly, the tuning of hyperparameters, namely, the clinical hyperparameter (EDP_{ref}) and the PPO hyperparameters (T, ϵ, b, and lr) and, secondly, the complete training to find optimal PPO parameters (network weights). The optimal EDP_{ref} is selected by training four different controllers at target values of 8, 12, 14, and 16 mmHg for 15 000 steps and later compared on an evaluation episode. These target values are in line with the study by Imamura et al.,[31] where the pulmonary capillary wedge pressure, closely related to EDP, is kept below 18 mmHg.

The optimal EDP_{ref} value should guarantee LV unloading but also a cardiac output (TCO) high enough to assure enough perfusion to end organs. Assuming a cardiac index of 2.5 and an average body surface of 2 m^2, a target TCO of 5.0 L/min is set, in line with the study by Imamura et al.[31]

The PPO hyperparameters are varied to produce five different controllers (see Table 3), which are also trained for 15 000 steps and evaluated based on their training performance. The *Baseline PPO* configuration in Table 3 is chosen based on the length of the training episode defined. For the second part of the training phase, to adjust network weights, the DRL controller is trained for 45 000 steps.

3 | RESULTS

3.1 | Training

In Figure 3, the evaluation performance of the controllers with a target EDP_{ref} of 8, 12, 14, and 16 mmHg are shown. Of the four controllers, the ones with a target EDP_{ref} of 14 and 16 mmHg assure a TCO the closest to the desired 5.0 L/min and the one with a target EDP_{ref} of 14 mmHg additionally shows more stability especially when imposing

Label	T	ϵ	b	lr (10^{-4})	Step (10^3)	Reward	Prop. Max. Reward
Baseline PPO	40	0.2	20	3	15	209	0.52
Horizon variation	20	0.2	20	3	15	223	0.55
Clip range variation	20	0.4	20	3	15	226	0.56
Batch size variation	**20**	**0.4**	**10**	**3**	**15**	**243**	**0.6**
Learning rate variation	20	0.4	10	30	15	229	0.57

[a] Horizon (T), clip range (ϵ), batch size (b), and learning rate (lr). The results were obtained with $EDP_{\text{ref}} = 14$ mmHg. In bold, the best configuration in terms of episodic mean reward of the last rollout.

TABLE 3 Episodic mean reward of last rollout at training step recorded for variations of multiple PPO hyperparameters.[a]

changes on *HR* while still allowing for consistent LV flow (see $Q_{LV\text{mean}}$ mean and *EDP* in Figure 3). This controller is retained for the further analysis described afterwards.

Different PPO configurations are compared based on their training performance. As shown in Table 3, the maximum episodic mean reward of the last rollout at step 15 000 is achieved with the *Batch size variation* configuration, which accounts for 60% of the maximum training episodic score of 402. Additionally, Figure 4 shows the evolution of the episodic reward during training.

3.2 | Evaluation

The controller with *Batch size variation* configuration from Table 3 is trained for 45 000 steps. From here on, the resulting controller is referred to as the Optimal DRL controller. Figure 5 compares the performance of the Optimal DRL controller and a LVAD running at 2400 rpm in the evaluation episode. This speed is chosen as the average of the Optimal DRL controller speed during the stabilization phases of the experiment parameters. The Optimal DRL Controller achieves a more stable *EDV* and *EDP* signal compared to the constant speed LVAD. This difference is greater during variations on *TBV*, *SVR*, *Ers*, and *Els*. Additionally, the Optimal DRL controller achieves a higher magnitude of aortic valve opening, allowing an increased LV flow, as seen in the subplots of $P_{LV\max} - P_{AO\max}$ and $Q_{LV\text{mean}}$ of Figure 5, while maintaining a similar *TCO* profile relative to the constant speed LVAD.

4 | DISCUSSION

The aim of this work is to develop an automatic, physiological control of the LVAD speed, able to dynamically respond to a patient's hemodynamic condition. We propose a preload-based DRL LVAD controller using a PPO algorithm built upon the data of a deterministic high-fidelity cardiorespiratory simulator. The Optimal DRL controller is able to react to changes in the cardiovascular system assuring an adequate perfusion to end organs, increasing flow through the aortic valve and increasing *EDV* stability, when compared to the constant speed LVAD strategy.

To avoid LVAD suction and promote aortic valve opening, a controller must be able to keep LV preload at an optimal level. The LV preload should be low enough to promote LV unloading and pulmonary decongestion, and high enough to allow the LV to contract and (if some residual contractility is left) open the aortic valve, according to the Frank–Starling mechanism.[3] Avoiding aortic valve closure can be more challenging, as there are more variables involved. This has been achieved by Amacher et al.,[32] using an optimal control strategy that aims to solve a sequential optimization problem based on a reduced model of the environment. In a study by Burkhoff et al.,[33] increases of the LVAD speed are shown to create a growing difference in pressure signals, which should overlap to allow for aortic valve opening.

The Optimal DRL controller is trained on different and clinically plausible states of a HF patient under LVAD support, produced with the cardiorespiratory simulator. For this purpose, specific changes of cardiovascular parameters are imposed on the simulator, of various entity and severity, and with a time scale compatible with the literature.[30] Changes in the cardiovascular system can be very dynamic due to different causes and circumstances such as medication, stress, physical activity, and fluid intake.[34,35] In addition, longer term changes can also be noticed specifically in the myocardial function, due, for example, to reverse remodeling of the heart after revascularization or unloading or to a worsening of the contractile function upon progression of disease.[36] We decided to change a set of six cardiovascular parameters (*TBV*, *HR*, *SVR*, *PVR*, *Els*, and *Ers*) to capture this large inter- and intra-patient variability observed in clinical practice and in patient's daily life. The change to these parameters is done sequentially one at a time, simplifying experimentation but, also, limiting generalization. Consequently, the effect that one parameter might have on another (e.g., ventricular interaction) is not modeled. Experiments were designed to capture a large patient

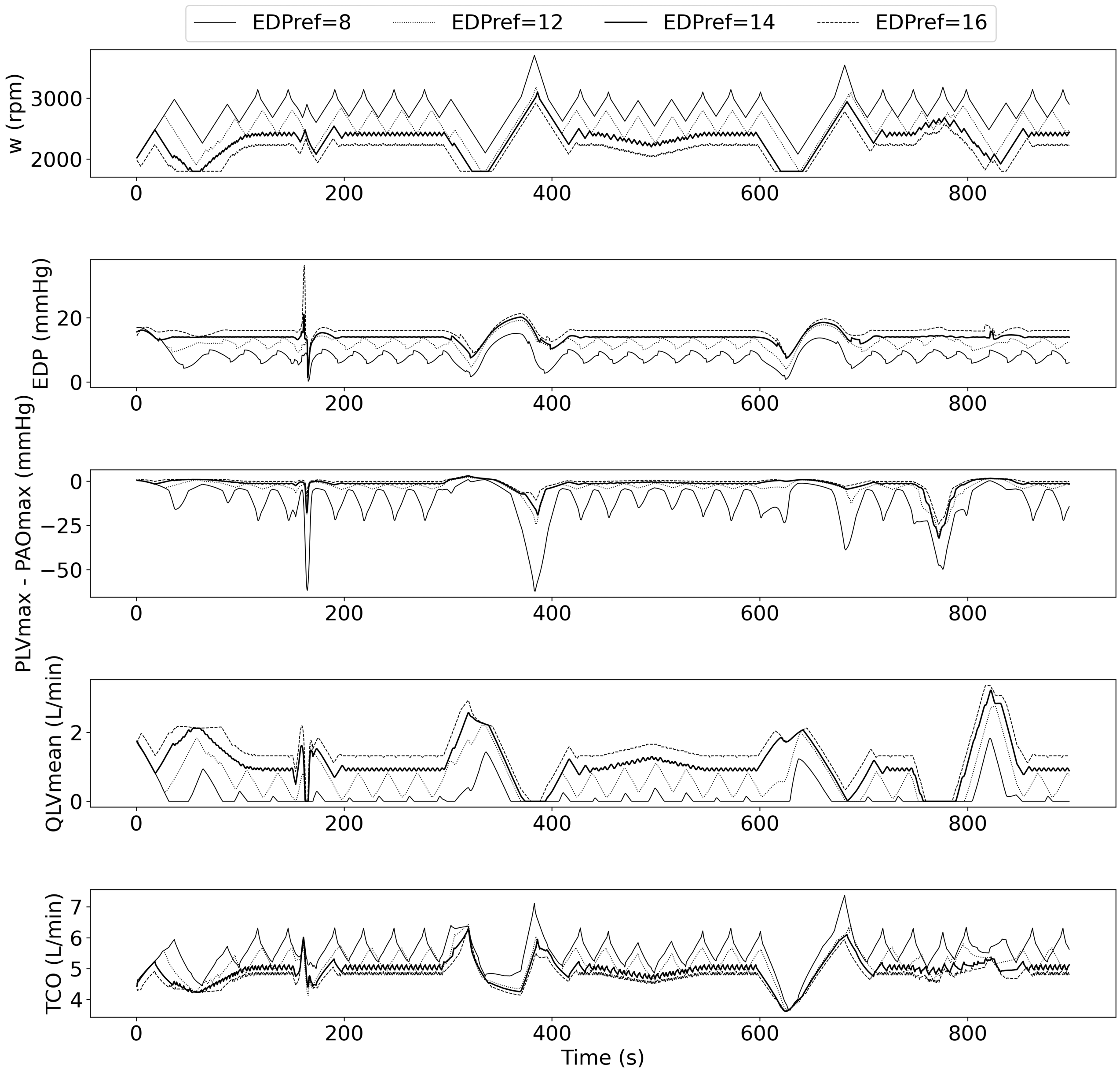

FIGURE 3 Influence of EDP_{ref} during the evaluation episode on: LVAD speed or ω (rpm), end-diastolic pressure or EDP (mmHg), $s_{\text{valve}} = P_{LV\text{max}} - P_{AO\text{max}}$ (mmHg), mean LV flow or $Q_{LV\text{mean}}$ (L/min), and total cardiac output or TCO (L/min). The results are obtained for a controller with the PPO hyperparameters set to the *Baseline PPO* configuration from Table 2.

variability. It could be argued that the ranges of change for these parameters could have been expanded even more, so to capture extreme conditions such as severe hypovolemia. While including more extreme situations to the experiments would yield a more robust controller, those situations requiring further medical intervention are out of the scope of this study. In Rocchi et al.,[37] authors outline a framework for dealing with such extreme conditions, such as severe hypovolemia, for which decreasing the LVAD speed is not enough and should be accompanied by additional medical interventions.

During the evaluation, LV suction is not occurring for the configured HF patient supported by the Optimal DRL controller and by the constant speed LVAD. Nevertheless, the Optimal DRL controller assures a more stable *EDV* with a standard deviation of 5 mL than the constant speed controller, which induces a *EDV* with a standard deviation of 9 mL. Furthermore, for the same patient profile, the Optimal DRL controller is able to achieve a higher degree of aortic flow producing an average $Q_{LV\text{mean}}$ of 1.1 L/min during evaluation, compared to the constant speed LVAD alternative, which delivers an average $Q_{LV\text{mean}}$ of 0.9 L/min during evaluation.

The Optimal DRL controller has two design strengths drawn from the work of other authors: having a LV preload-based design and depending on inputs based

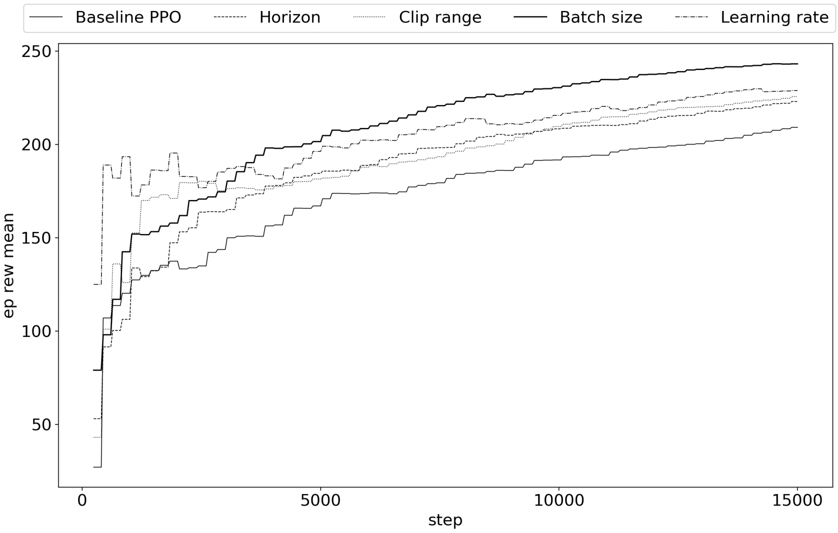

FIGURE 4 Evolution of the mean episodic reward by step, for all configurations of the PPO agent from Table 3.

on pressure signals. First and foremost, the proposed design is preload-based in line with previous work,[38–40] which is motivated by the fact that ventricles are responsive to preload variations from a hemodynamic point of view.[3] Having a preload-based controller for the LVAD is an important prerequisite for a physiological controller, but it is also important to consider the feasibility of measuring the variables involved. Secondly, the Optimal DRL controller depends on LV pressure signals as inputs, which are more feasible to be measured continuously than volume signals. In a study by Ochsner et al.,[7] a pressure-based controller,[41] a volume-based controller,[39] and a constant speed controller are compared in an in vivo trial. Results of the comparison show that the pressure-based controller performed the best of all three, indicating that the volume sensors required are a design weakness in LVAD controllers, due to the complications they introduce. Implantable pressure sensors have previously been used as a long-term solution to input pressure signals into LVAD controllers.[42]

Not much research has been done in DRL controllers for LVAD. Recent work by Li et al.[9] presents a DRL controller based on a soft-actor-critic RL agent that is both quicker and more effective in keeping an adequate LVAD flow to the patient, with respect to a theoretical target pump flow. The results are validated in a large experiment that randomizes multiple predefined parameters to create inter- and intra-patient variability. The experiments include testing perturbations on single (e.g., systemic blood resistance) and mixed (systemic blood resistance, myocardial contractile force and heart rate) factors. However, the controller by Li et al.[9] has two main limitations. Firstly, the design depends on a theoretical target pump flow. While this is helpful in designing the RL controller by avoiding the sparsity problem (having a distance based term in the states and the rewards), it can also lead to a lack of generalization for a wide enough range of patients, which might not be captured in the experiments. The reason is that the theoretical target pump flow used by Li et al.[9] is based on a general physiological relationship between pump flow and ventricular preload, without accounting for subject variability in the patient population.[43,44] In contrast, our Optimal DRL controller avoids relying on a theoretical target pump flow by tuning a clinical parameter $\left(EDP_{\text{ref}}\right)$ during an evaluation of the controller in an extensive experiment of six cardiovascular parameters (TBV, HR, SVR, PVR, Els, Ers). Secondly, in the work by Li et al.,[9] the DRL controller is trained and tested on a relatively simple cardiovascular simulator, which accounts for the LV and the systemic circulation, leaving out from the simulator the right ventricle and the pulmonary circulation. Conversely, for our work, a high-fidelity cardiorespiratory simulator is used that allows to produce large data samples with a wide range of inter- and intra-patient variability, which is key when building a robust DRL controller.

Despite the positive results, the Optimal DRL controller presents three main limitations. Firstly, the Optimal DRL controller underperforms when tested against HR changes, as it fails to react quickly enough to prevent large variations of EDV and EDP. This is due to the discretization of the action space of the controller. This shortcoming is overcome in the work by Li et al.[9] by having an adaptive effective action space to quickly reach desired LVAD speeds. Secondly, the Optimal DRL controller includes actions in the reward function. This constrains the learning possibilities of the agent as the optimal strategy is explicit in the reward. However, this design choice led to an improvement in training efficiency and has been previously implemented in RL controllers for AP.[11] Thirdly, the Optimal DRL controller relies on tuning EDP_{ref}. This, instead, could be part

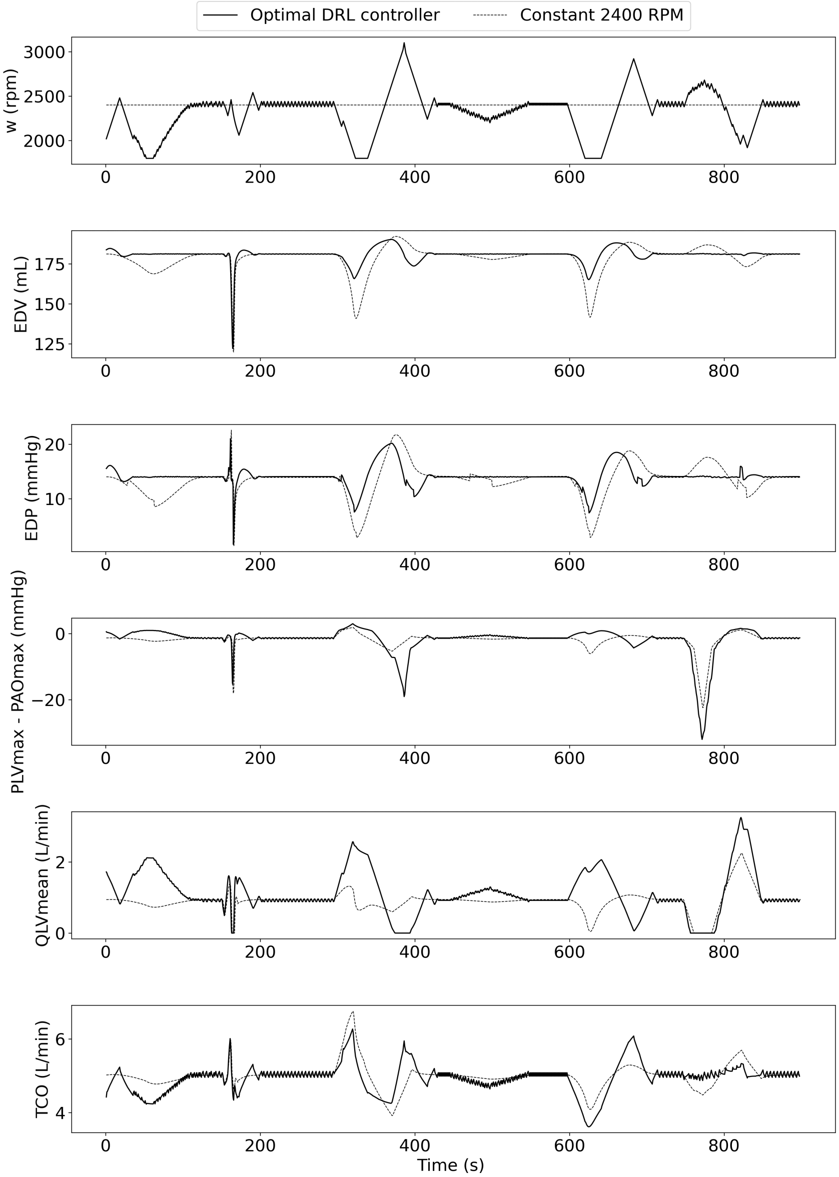

FIGURE 5 Comparison between the performance of the Optimal DRL controller and a constant LVAD speed of 2400 rpm, in terms of: LVAD speed or ω (rpm), end-diastolic volume or *EDV* (mL), end-diastolic pressure or *EDP* (mmHg), $s_{\text{valve}} = P_{LV\max} - P_{AO\max}$ (mmHg), mean LV flow or $Q_{LV\text{mean}}$ (L/min), and total cardiac output or *TCO* (L/min).

of the action space of the controller, achieving a greater applicability to patient profiles that require a different EDP_{ref} to the one selected in this work. Implementing this improvement requires adding a second dimension to the action space: the controller would act on ω and EDP_{ref}. The new reward function would be composed of an additional third term to the previous reward function. In this term, the implied goal would be to produce a stable positive s_{valve} signal by making moderate changes on EDP_{ref}, following the results of Figure 3, which shows that increases in EDP_{ref} promote aortic valve opening.

5 | CONCLUSION

This article introduces the Optimal DRL controller based on a PPO agent trained with data from a deterministic high-fidelity cardiorespiratory simulator, which introduces the possibility to create realistic environments to train and test RL-based LVAD controllers. In this work, the controller is trained and tested on an extensive set of experiments realized by changing six cardiovascular parameters, to realistically capture intra- and inter-patient variability. The Optimal DRL controller we obtained performs better than the constant speed LVAD, increasing flow through the aortic valve and increasing *EDV* stability, when compared to the constant speed LVAD strategy.

AUTHOR CONTRIBUTIONS

Diego Fernández-Zapico: conceptualization, methodology, formal analysis, investigation, software, writing, visualization. **Thijs Peirelinck**: conceptualization, formal analysis, methodology, writing (review and editing), critical review, supervision. **Geert Deconinck**: critical revision and approval. **Dirk W. Donker**: critical revision and approval. **Libera Fresiello**: conceptualization, formal analysis, investigation, software, writing (review and editing), critical review, supervision, resources.

FUNDING INFORMATION

None reported.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

ORCID

Diego Fernández-Zapico https://orcid.org/0009-0004-7883-4221
Thijs Peirelinck https://orcid.org/0000-0002-1080-8164
Geert Deconinck https://orcid.org/0000-0002-2225-3987

REFERENCES

1. WHO. The top 10 causes of death 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. Bragazzi NL, Zhong W, Shu J, Abu Much A, Lotan D, Grupper A, et al. Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017. Eur J Prev Cardiol. 2021;28(15):1682–90. https://doi.org/10.1093/eurjpc/zwaa147
3. Karimov J, Fukamachi K, Starling R. Mechanical support for heart failure. Cham, Switzerland: Springer; 2020. https://doi.org/10.1007/978-3-030-47809-4
4. Long B, Robertson J, Koyfman A, Brady W. Left ventricular assist devices and their complications: a review for emergency clinicians. The. Am J Emerg Med. 2019;37(8):1562–70. https://doi.org/10.1016/j.ajem.2019.04.050
5. Stevens MC, Stephens A, AlOmari AHH, Moscato F. Chapter 20 – physiological control. In: Gregory SD, Stevens MC, Fraser JF, editors. Mechanical circulatory and respiratory support. London, UK: Academic Press; 2018. https://doi.org/10.1016/B978-0-12-810491-0.00020-5
6. Stöcklmayer C, Dorffner G, Schmidt C, Schima H. An artificial neural network-based noninvasive detector for suction and left atrium pressure in the control of rotary blood pumps: an in vitro study. Artif Organs. 1995;19(7):719–24. https://doi.org/10.1111/j.1525-1594.1995.tb02411.x
7. Ochsner G, Wilhelm MJ, Amacher R, Petrou A, Cesarovic N, Staufert S, et al. In vivo evaluation of physiologic control algorithms for left ventricular assist devices based on left ventricular volume or pressure. ASAIO J. 2017;63(5):568–77. https://doi.org/10.1097/MAT.0000000000000533
8. Maw M, Schlöglhofer T, Marko C, Aigner P, Gross C, Widhalm G, et al. A Sensorless modular multiobjective control algorithm for left ventricular assist devices: a clinical pilot study. Front Cardiovasc Med. 2022;9. https://doi.org/10.3389/fcvm.2022.888269
9. Li T, Cui W, Xie N, Li H, Liu H, Li X, et al. Intelligent and strong robust CVS-LVAD control based on soft-actor-critic algorithm. Artif Intell Med. 2022;128:1023. https://doi.org/10.1016/j.artmed.2022.102308
10. Coronato A, Naeem M, De Pietro G, Paragliola G. Reinforcement learning for intelligent healthcare applications: a survey. Artif Intell Med. 2020;109:101964. https://doi.org/10.1016/j.artmed.2020.101964
11. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. Artif Intell Med. 2020;104. https://doi.org/10.1016/j.artmed.2020.101836
12. Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Rev Med Devices. 2013;10(5):661–73. https://doi.org/10.1586/17434440.2013.827515
13. Man CD, Micheletto F, Lv D, Breton M, Kovatchev B, Cobelli C. The UVA/PADOVA type 1 diabetes simulator: new features. J Diabetes Sci Technol. 2014;8(1):26–34. https://doi.org/10.1177/1932296813514502
14. Wilson C, Riccardi A. Improving the efficiency of reinforcement learning for a spacecraft powered descent with Q-learning. Optim Eng. 2023;24:223–55. https://doi.org/10.1007/s11081-021-09687-z
15. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. 33rd conference on uncertainty in artificial intelligence; 2017. Available from: http://www.scopus.com/inward/record.url?scp=85031120880&partnerID=8YFLogxK
16. Padmanabhan R, Meskin N, Haddad WM. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. Biomed Signal Process Control. 2015;22:54–64. https://doi.org/10.1016/j.bspc.2015.05.013
17. Fetanat M, Stevens M, Hayward C, Lovell NH. A Sensorless control system for an implantable heart pump using a real-time deep convolutional neural network. IEEE Trans Biomed Eng. 2021;68(10):3029–38. https://doi.org/10.1109/TBME.2021.3061405
18. Ng BC, Salamonsen RF, Gregory SD, Stevens MC, Wu Y, Mansouri M, et al. Application of multiobjective neural predictive control to biventricular assistance using dual rotary blood pumps. Biomed Signal Process Control. 2018;39:81–93. https://doi.org/10.1016/j.bspc.2017.07.028
19. Mahr C, Chivukula VK, McGah P, Prisco AR, Beckman JA, Mokadam NA, et al. Intermittent aortic valve opening and risk of thrombosis in ventricular assist device patients. ASAIO J. 2017;63(4):425–32. https://doi.org/10.1097/MAT.0000000000000512
20. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, USA: MIT Press; 2018.
21. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In: Bach F, Blei D, editors. Proceedings of the 32nd international conference on machine learning. Vol. 37 of proceedings of machine learning research. Lille, France: PMLR; 2015. Available from: https://proceedings.mlr.press/v37/schulman15.html
22. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. https://doi.org/10.48550/arXiv.1707.06347
23. Wiering M, van Otterlo M. Reinforcement learning: state-of-the-art. Heidelberg, Germany: Springer; 2012. https://doi.org/10.1007/978-3-642-27645-3

24. Stable Baselines3. Stable-baselines3 docs – reliable reinforcement learning implementations 2022. Available from: https://stable-baselines3.readthedocs.io/en/master/
25. Fresiello L, Rademakers F, Claus P, Ferrari G, Di Molfetta A, Meyns B. Exercise physiology with a left ventricular assist device: analysis of heart-pump interaction with a computational simulator. PLoS One. 2017;12(7). https://doi.org/10.1371/journal.pone.0181879
26. Fresiello L, Meyns B, Di Molfetta A, Ferrari G. A model of the cardiorespiratory response to aerobic exercise in healthy and heart failure conditions. Front Physiol. 2016;7. https://doi.org/10.3389/fphys.2016.00189
27. Fresiello L, Ferrari G, Di Molfetta A, Zieliński K, Tzallas A, Jacobs S, et al. A cardiovascular simulator tailored for training and clinical uses. J Biomed Inform. 2015;57:100–12. https://doi.org/10.1016/j.jbi.2015.07.004
28. Gym. Gym documentation 2022. Available from: https://www.gymlibrary.dev/
29. Fresiello L, Gross C, Jacobs S. Exercise physiology in left ventricular assist device patients: insights from hemodynamic simulations. Ann Cardiothorac Surg. 2021;10(3). https://doi.org/10.21037/acs-2020-cfmcs-23
30. Ursino M. Interaction between carotid baroregulation and the pulsating heart: a mathematical model. Am J Physiol Heart Circ Phys Ther. 1998;275(5):1733–47. https://doi.org/10.1152/ajpheart.1998.275.5.H1733
31. Imamura T, Jeevanandam V, Kim G, Raikhelkar J, Sarswat N, Kalantari S, et al. Optimal hemodynamics during left ventricular assist device support are associated with reduced readmission rates. Circ Heart Fail. 2019;12(2). https://doi.org/10.1161/CIRCHEARTFAILURE.118.005094
32. Amacher R, Asprion J, Ochsner G, Tevaearai H, Wilhelm MJ, Plass A, et al. Numerical optimal control of turbo dynamic ventricular assist devices. Bioengineering. 2014;1(1):22–46. https://doi.org/10.3390/bioengineering1010022
33. Burkhoff D, Sayer G, Doshi D, Uriel N. Hemodynamics of mechanical circulatory support. J Am Coll Cardiol. 2015;66:2663–74. https://doi.org/10.1016/j.jacc.2015.10.017
34. Fresiello L, Buys R, Timmermans P, Vandersmissen K, Jacobs S, Droogne W, et al. Exercise capacity in ventricular assist device patients: clinical relevance of pump speed and power. Eur J Cardiothorac Surg. 2016;50(4):752–7. https://doi.org/10.1093/ejcts/ezw147
35. Imamura T, Chung B, Nguyen A, Sayer G, Uriel N. Clinical implications of hemodynamic assessment during left ventricular assist device therapy. J Cardiol. 2018;71(4):352–8. https://doi.org/10.1016/j.jjcc.2017.12.001
36. Marinescu KK, Uriel N, Mann DL, Burkhoff D. Left ventricular assist device-induced reverse remodeling: it's not just about myocardial recovery. Expert Rev Med Devices. 2017;14(1):15–26. https://doi.org/10.1080/17434440.2017.1262762
37. Rocchi M, Libera FS, Jacobs DD, Droogne W, Meyns B. Potential of medical management to mitigate suction events in ventricular assist device patients. ASAIO J. 2022;68(6):814–21. https://doi.org/10.1097/MAT.0000000000001573
38. Bullister E, Reich S, Sluetz J. Physiologic control algorithms for rotary blood pumps using pressure sensor input. Artif Organs. 2002;26(11):931–8. https://doi.org/10.1046/j.1525-1594.2002.07126.x
39. Ochsner G, Amacher R, Wilhelm MJ, Vandenberghe S, Tevaearai H, Plass A, et al. A physiological controller for turbodynamic ventricular assist devices based on a measurement of the left ventricular volume. Artif Organs. 2014;38(7):527–38. https://doi.org/10.1111/aor.12225
40. Mansouri M, Salamonsen RF, Lim E, Akmeliawati R, Lovell NH. Preload-based Starling-like control for rotary blood pumps: numerical comparison with pulsatility control and constant speed operation. PLoS One. 2015;10(4). https://doi.org/10.1371/journal.pone.0121413
41. Petrou A, Ochsner G, Amacher R, Pergantis P, Rebholz M, Meboldt M, et al. A physiological controller for turbodynamic ventricular assist devices based on left ventricular systolic pressure. Artif Organs. 2016;40(9):842–55. https://doi.org/10.1111/aor.12820
42. Dual SA, Zimmermann JM, Neuenschwander J, Cohrs NH, Solowjowa N, Stark WJ, et al. Ultrasonic sensor concept to fit a ventricular assist device cannula evaluated using geometrically accurate heart phantoms. Artif Organs. 2019;43(5):467–77. https://doi.org/10.1111/aor.13379
43. Stevens MC, Gaddum NR, Pearcy M, Salamonsen RF, Timms DL, Mason DG, et al. Frank-starling control of a left ventricular assist device. Annual international conference of the IEEE engineering in medicine and biology society; 2011; 2011. https://doi.org/10.1109/IEMBS.2011.6090314
44. Hall JE. Chapter 9 – cardiac muscle; the heart as a pump and function of the heart valves. Guyton and Hall textbook of medical physiology. 14th ed. Philadelphia, USA: Elsevier; 2021.

How to cite this article: Fernández-Zapico D, Peirelinck T, Deconinck G, Donker DW, Fresiello L. Physiological control for left ventricular assist devices based on deep reinforcement learning. Artif. Organs. 2024;48:1418–1429. https://doi.org/10.1111/aor.14845