POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Exploring the Potential of Reinforcement Learning for Automated Anesthesia Control

Laurea Magistrale in Computer Engineering - Ingegneria Informatica

Author: Manuela Merlo

Advisor: Prof. Francesco Trovò

Co-advisors: Prof. Alberto Maria Metelli, Dr. Emiliano Tognoli

Academic year: 2023-2024

1. Introduction

Each year, more than 300 million procedures worldwide require general anesthesia, a medically induced reversible state that enables patients to undergo surgery safely. Currently, anesthesiologists rely on monitoring systems to assess patient state and adjust drug administration accordingly. However, despite significant advances, considerable inter- and intra-patient variability makes it difficult to determine the right drug dosages. Additionally, human factors increase the likelihood of making mistakes. Errors in anesthetic drug administration can cause intraoperative awareness, adverse outcomes, and even mortality. This explains the need for controllers to optimize delivery strategies. Automated systems have been shown to improve patient safety. Among the most common approaches are Proportional-Integral-Derivative (PID) controllers and Model Predictive Control (MPC). While both offer proven benefits, they have limitations, such as reduced adaptability due to fixed parameters or reliance on an accurate model. Reinforcement Learning (RL) [5] offers a model-free and adaptive approach for patient-specific treatments. Studies have demonstrated RL's ability to control anesthesia effectively. However, existing implementations have mostly focused on regulating a single anesthetic agent and/or controlling a single physiologic condition rather than jointly managing multiple drugs and parameters. These systems disregard the synergy between anesthetics, reducing the effectiveness of automated control. This work presents three RL agents, all aimed at managing the maintenance phase of anesthesia without accounting for changes in the patient's condition due to external stimuli or complications. The controllers regulate propofol, a hypnotic, and remifentanil, an analgesic, commonly used for closed-loop anesthesia control. The first agent shares the same objective of a PID controller designed to optimize propofol for regulating Depth of Hypnosis (DoH) using a fixed propofol-to-remifentanil ratio. The second agent incorporates patient demographic data and removes the fixed ratio assumption. The third adds hemodynamic stability regulation. Despite its potential, RL needs large volumes of high-quality training data that are not easily accessible. Some studies not relying on clinical settings use simple models or simulators to design and evaluate decision-making systems. However, simulators in the literature

do not use physiological models and consider a linear interaction between drug responses to predict hemodynamic variables. We address this challenge by introducing the Anesthesia Response Simulator (AReS). This tool models patients' responses to common anesthetic drugs and shows its consistency with real anesthesia measurements. Using a simulator, a RL agent learns from Pharmacokinetic-Pharmacodynamic models, like those used in MPC. The advantage is that RL can generalize beyond a fixed model. However, its real-world effectiveness depends on the simulator's accuracy. This work serves as a **proof of concept**, demonstrating the feasibility of using RL with a complex model in simulation. We use SHapley Additive exPlanations (SHAP) [2] values to demonstrate that RL agents can exploit demographic data to make the therapy patient-specific. The next step would be to collect real-world data to assess existing strategies and leverage RL to discover complex patterns for personalized treatments that go beyond traditional models.
In summary, the main contributions of this thesis are as follows:

- Development of a novel patient anesthesia simulator that provides an in-simulation environment for training and testing anesthesia controllers.
- Implementation of three RL-based controllers.
- Comparison of simulator results with clinical data.
- Performance evaluation of RL against a PID controller using standard metrics.
- Analysis of RL-based decision-making using explainable techniques.

2. Medical Background

General Anesthesia (GA) is a medically induced, reversible state that allows patients to undergo surgery. It involves the administration of hypnotics, opioids, and neuromuscular blocking agents to induce amnesia, hypnosis, analgesia, and akinesia. Specifically, propofol, remifentanil, and rocuronium are often used for their fast onset and recovery. Anesthesia consists of three phases. During *induction*, a bolus of a hypnotic agent is used to transition the patient from consciousness to unconsciousness. *Maintenance* sustains anesthesia with a continuous infusion of hypnotics and opioids. Finally, drug administration is discontinued during *emergence* to restore patient consciousness. Anesthesiologists use the Bispectral Index (BIS) to evaluate hypnosis depth. However, due to monitors' delay in displaying the signal, a new index called NeuroWave Index (WAV_{cns}) has been developed to provide real-time measurements. A value in the $[40, 60]$ range is considered safe to prevent awareness and poor outcomes. Mean Arterial Pressure (MAP) is used to assess organ dysfunction risks. In addition, in the case of hypotension, Cardiac Output (CO), Stroke Volume (SV), and Heart Rate (HR) help to identify its cause. Guidelines recommend a MAP target of at least 65 mmHg. Finally, Neuromuscular Blockade (NMB) is monitored using Train-Of-Four (TOF) stimulation or Post-Tetanic Count (PTC) depending on the level of akinesia required. Both techniques use electric stimuli and assess the amplitude or number of contractions. Clinicians rely on Pharmacokinetic (PK) models that describe the evolution of drug concentration in the body and Pharmacodynamic (PD) models that relate concentrations to the drug's therapeutic effects. Target-Controlled Infusion (TCI) systems currently help anesthesiologists administer anesthesia by computing the required drug dose to achieve a user-defined concentration in the patient based on PK-PD models. However, since these models might be inaccurate, the patient reaction may differ from the one the user desires. Although advancements have been made, anesthesia administration remains challenging because of the risks of underdosing and overdosing. Indeed, insufficient anesthesia can cause intraoperative awareness, while excessive doses are linked to poor outcomes and increased mortality. Inter- and intra-patient variability, influenced by factors such as age, weight, sex, and blood loss, makes it difficult to define the appropriate amount of drugs to be administered for adequate anesthesia. In addition, human errors caused by inexperience, fatigue, and distraction further contribute to patient safety risks.

3. Problem Formulation

This work focuses on managing the maintenance phase of anesthesia without accounting for changes in the patient's condition due to

external stimuli or complications. We present the task as a progressive learning framework, with objectives of increasing complexity. Each problem is expressed as a Markov Decision Process (MDP) [4], defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ represents the transition dynamics, $R(s,a)$ is the reward function, and $\gamma \in [0,1]$ is the discount factor.

We propose three RL agents. The first agent controls the DoH using a fixed propofol-to-remifentanil ratio. The state space consists of the proportional, integral, and derivative errors of the WAV_{cns} index relative to a target of 50. The action space is the propofol infusion rate, and the reward is defined as:

$$R = -|\text{WAV}_{\text{error}}| = -\left|\frac{\text{WAV} - 50}{50}\right|.$$

The second agent removes the fixed drug ratio constraint, allowing free control of propofol and remifentanil. The state space includes the patient demographics (age, height, weight, gender) and the $\text{WAV}_{\text{error}}$. The reward function introduces a penalty term to minimize drug use:

$$R = -|\text{WAV}_{\text{error}}| - 0.5\ (a^2_{\text{prop}} + a^2_{\text{remi}}).$$

The third agent extends control to hemodynamic stability by including the MAP error in the state space. A dead zone is introduced in the $[65, 75]$ mmHg range, where no penalty is applied. Outside this range, the MAP error is defined as:

$$\text{MAP}_{\text{error}} = \begin{cases} \frac{65-\text{MAP}}{65}, & \text{if MAP} < 65 \\ \frac{\text{MAP}-75}{75}, & \text{if MAP} > 75 \end{cases}.$$

The final reward function extends the previous formulation by incorporating $\text{MAP}_{\text{error}}$:

$$R = -|\text{WAV}_{\text{error}}| - \text{MAP}_{\text{error}} - 0.5\ (a^2_{\text{prop}} + a^2_{\text{remi}}).$$

We set all the discount factors to 0.99.

4. Simulator Development

Training RL agents for anesthesia automation requires large datasets that explore a wide range of control actions and physiological states. Due to ethical concerns and patient safety risks, such clinical data is not easily accessible. To overcome this limitation, we developed the Anesthesia Response Simulator to mimic the response of patients undergoing anesthesia. As shown in Figure 1, the tool consists of five components: 1) the **input** unit to indicate drug dosages, 2) **Pharmacokinetic** models to calculate plasma concentrations, 3) **Pharmacodynamic** models to compute therapeutic effects, 4) the **disturbance** model to emulate the impact of surgical stimuli and blood loss, and 5) the **output** unit to access states predictions. The input unit receives the infusion rates or target concentrations of propofol, remifentanil, norepinephrine, and rocuronium. The simulator's outputs include BIS and WAV_{cns} to assess DoH, and MAP, CO, HR, and SV to evaluate hemodynamic stability. PTC is used to monitor profound NMB.

The simulator employs population-based PK models that consider patient-specific covariates to predict plasma concentrations of propofol, remifentanil, norepinephrine, and rocuronium. Regarding PD modeling, AReS allows users to simulate DoH, considering propofol and remifentanil plasma concentrations. Depending on the user-selected interaction mode, the simulator can compute the effect using only propofol or account for the synergy between propofol and remifentanil. Hemodynamic variables are computed using a mechanism-based PD model incorporating the General Pharmacodynamic Interaction (GPDI) framework to represent the non-linear effects of hypnotics and opioids on Total Peripheral Resistance (TPR), HR, and SV. MAP is computed as the product of CO and TPR and CO as the product of HR and SV. The effect of norepinephrine on MAP follows a standard nonlinear model. Its influence on CO is estimated based on observations taken from the literature. Since the interaction between propofol, remifentanil, and norepinephrine has not been previously modeled, we assume a linear interaction between these drugs affects hemodynamic variables. Finally, the NMB PD model outputs the vector $(p_{\text{moderate}}(t),\ p_{\text{deep}}(t),\ p_{\text{profound}}(t),\ p_{\text{very profound}}(t))$ where $p_m(t)$ describes the probability of observing category m.

In addition to drug effects, AReS also models patients' reactions to surgical stimuli and hypovolemia. The stress response is represented as an additive value applied to DoH, HR, and TPR. The magnitude of the response depends on the user-defined maximum and the patient's

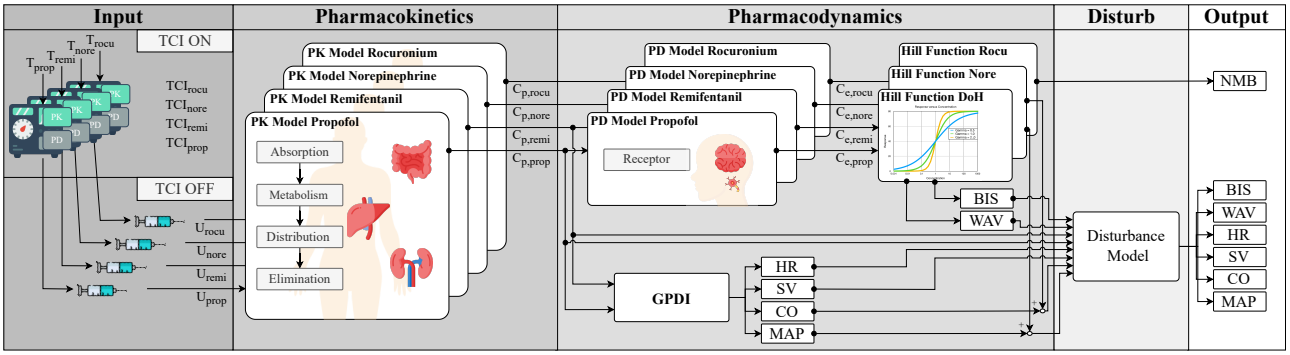

Figure 1: Schematic view of the simulator.

drug concentrations, with higher anesthetic levels reducing it. Finally, the simulator represents hypovolemic states by modifying hemodynamic variables with multiplicative coefficients, the values of which are taken from the literature.

5. RL Framework

In the following, we describe the simulator configuration used in the environment we developed to train RL agents and provide an overview of the employed algorithm, explaining the rationale behind our decision. Finally, we outline the hyperparameters and architectural choices.

5.1. Environment

We allow agents to control the infusion rates of propofol and remifentanil. Each episode lasts $T_{sim} = 5000$ seconds, with actions issued every 5 seconds. Agents interact with different patients during each episode, whose demographic data are randomly generated. We use the Eleveld PK and PD models for both drugs as they can represent a diverse patient population. Additionally, we account for drugs' synergistic effects.

5.2. Learning Algorithm

The three agents are trained using the Twin Delayed Deep Deterministic Policy Gradient (TD3) [1] algorithm, chosen for its compatibility with continuous action spaces and deterministic policy, enhancing trust in health applications. This model-free online algorithm follows an actor-critic structure in which the actor learns the policy, and the critic evaluates it. In particular, the TD3 algorithm improves Deep Deterministic Policy Gradient (DDPG), addressing overestimation bias by integrating two critics and stabilizing training with delayed policy updates.

5.3. Hyperparameters and Architectural Design

All agents share a common architecture where the actor and critic networks are fully connected neural networks. Each network has two hidden layers, with 64 neurons in the first layer and 32 in the second, using ReLU activation functions. In addition, the actor's output layer includes a Tanh activation function. The learning rate is fixed at 0.001, and Gaussian noise is added to actions for exploration, with a decay rate specific to each agent and action. The chosen optimizer is Adam. Other key hyperparameters include a replay buffer size of 1 M and a batch size of 256. The key difference between agents lies in their feature extraction mechanisms. Agent 1 employs a simple FlattenExtractor that takes the input tensor and flattens it into a 1D vector. In contrast, Agents 2 and 3 employ an approach inspired by Feature-wise Linear Modulation (FiLM) [3] to transform feature representations based on conditioning variables.
In the second agent, a Neural Network (NN) takes patient demographics (age, height, weight, and sex) to generate a conditioning context. Another NN then uses this context to produce the scale and bias terms, which modulate the $\text{WAV}_{\text{error}}$ feature through an affine transformation:

$$\widehat{\text{WAV}}_{\text{error}} = \gamma_1 \cdot \text{WAV}_{\text{error}} + \beta_1.$$

The conditioned $\widehat{\text{WAV}}_{\text{error}}$, raw $\text{WAV}_{\text{error}}$, and demographic context are then processed by the last combining NN. The third agent's feature extractor in Figure 2 builds on this approach by

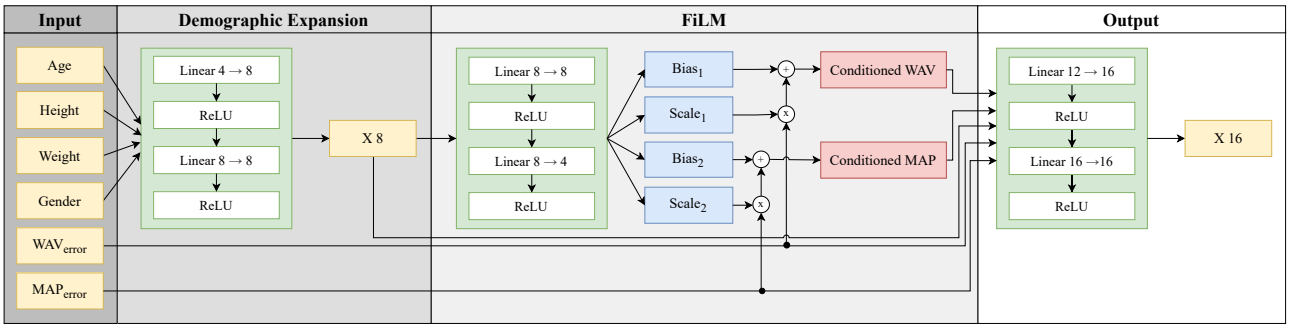

Figure 2: Feature Extractor Agent 3.

introducing an additional conditioning step for the MAP_{error}. The final vector integrates both raw and conditioned WAV_{error}/MAP_{error} values, along with the demographic context.

6. Evaluation

In this section, we validate AReS by comparing its output with real data. In addition, we evaluate the agents' effectiveness using standard control metrics and apply explainability methods to obtain information about their decision-making.

6.1. AReS Evaluation

Using recorded infusion rate values from a dataset, we evaluate whether the simulator outputs align with the actual measurements taken during anesthesia. We select data that include both WAV_{cns} and hemodynamic variables and configure the simulator to match the conditions under which the signals were recorded. The computed and measured signals are then compared using the Median Absolute Error (MedAE) as a performance metric. The results show that the simulator effectively captures the physiological effects of propofol and remifentanil at a population level, making it suitable for designing an anesthesia controller. However, inter- and intra-patient variability restricts its ability to make perfect individual predictions.

6.2. Controllers Evaluation

We evaluate the three agents with and without disturbances, using demographic data from 44 real patients. We implement a PID controller to establish a baseline. The parameters of the PID are optimized to adjust the DoH using a fixed propofol-to-remifentanil ratio via a genetic algorithm, minimizing the Integral of Absolute Error (IAE) with respect to the WAV_{error} across simulations of the same 44 patients. Patient response is simulated using a configuration consistent with the one used during agent training.
Their performance is measured using the Median Performance Error (MDPE, bias), the Median Absolute Performance Error (MdAPE, accuracy), the Wobble (stability), and the Root Mean Square Error (RMSE, error magnitude). Table 1 summarizes the results with metrics calculated against target values of 50 and 70 for WAV_{cns} and MAP, respectively. Each patient's errors are computed individually and then presented as mean ± standard deviation.
All controllers maintain DoH considerably within acceptable clinical thresholds. The first agent performs better than the PID by achieving a similar control strategy with better performance, reducing error while using the exact amount of drugs. The second and third agents achieve comparable performance despite their more complex objectives. Indeed, if we consider the averaged MdAPE values for the PID, the second and third agents are 0.38, 0.73, and 1.07. For a target value of 50, the controlled WAV_{cns} ranges from 49.81 to 50.19, 49.46 to 50.54, and 49.27 to 50.73, respectively, which are negligible for assessing DoH. Moreover, the third agent can balance DoH and hemodynamic stability. The agents also demonstrate robustness to surgical stimuli, even though they are not explicitly trained to handle them. However, while they can effectively compensate for perturbations in WAV_{cns}, they struggle with MAP. This is to be expected because the first two agents do not consider hemodynamic stability, and even the third agent, despite including it in its objective, is not exposed to extreme changes in the MAP during training.

Table 1: Performance metrics for PID and RL agents under disturbance-free conditions.

	Metric	PID	Agent 1	Agent 2	Agent 3	Threshold
WAV_{cns}	MDPE (%)	-0.24 ± 0.29	**-0.10 ± 0.17**	-0.20 ± 0.75	-0.87 ± 0.41	± 5.0
	MDAPE (%)	0.38 ± 0.16	**0.10 ± 0.17**	0.73 ± 0.29	1.07 ± 0.19	7.5
	Wobble (%)	0.20 ± 0.16	**0.03 ± 0.05**	0.09 ± 0.04	0.74 ± 0.42	5.0
	Controlled ±5 (%)	99.92 ± 0.51	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	≥ 80
	RMSE	0.51 ± 0.35	**0.16 ± 0.05**	0.37 ± 0.14	0.65 ± 0.07	5.0
MAP	MDPE (%)	-2.62 ± 3.58	-2.59 ± 3.59	-1.39 ± 3.59	**0.94 ± 4.25**	± 5.0
	MDAPE (%)	3.34 ± 3.14	3.13 ± 3.13	**2.72 ± 2.72**	3.70 ± 2.32	7.5
	Wobble (%)	**0.17 ± 0.07**	0.19 ± 0.07	0.29 ± 0.07	0.51 ± 0.06	5.0
	Controlled ±5 (%)	86.32 ± 33.98	86.36 ± 34.31	88.08 ± 31.75	**93.05 ± 21.00**	≥ 80
	RMSE	2.22 ± 2.22	2.24 ± 2.17	**2.02 ± 1.89**	2.60 ± 1.46	5.0

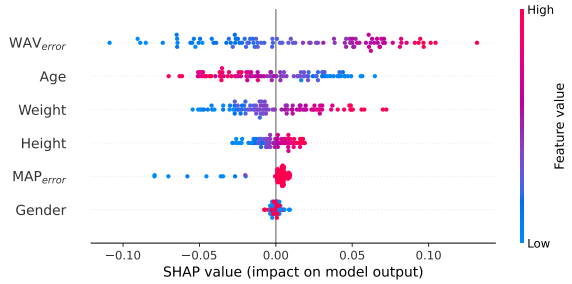

Figure 3: SHAP values propofol action Agent 3.

6.3. Explaining RL Agents Behavior

This section investigates the explainability of RL agents. We interpret how patient demographics influence the agents' decision-making employing SHAP [2] analysis, specifically the ExactExplainer from the SHAP library. An important result is that Agents 2 and 3 learned policies that align with the PK-PD models they were trained on. Figure 3 illustrates the shape values for the propofol action for the third agent. Here, the shift of dots to the right or left of the 0 value on the x-axis indicates the corresponding increase or decrease in the action due to changes in the input feature, relative to the expected model prediction when no features are considered. Specifically, Agent 3 increases propofol administration for positive WAV_{error} values, which correspond to a higher patient awareness. Additionally, it learned to reduce the dosage for elderly patients and increase it for heavier ones, following Eleveld's studies. Interpreting the influence of height is complex due to its relationship with weight. Finally, negative map errors cause the agent to reduce hypnotic administration, while gender has a minimal impact on decision-making. This demonstrates that RL agents successfully capture the relationships between demographics, drug administration, and physiological responses. More importantly, it highlights RL's potential to learn complex, real-world physiological patterns beyond the simulator, paving the way for personalized, patient-specific anesthesia control.

7. Conclusion

This thesis demonstrates the applicability of Reinforcement Learning in anesthesia control, showing that RL agents can learn policies that align with established PK and PD models. However, despite these promising results, several limitations must be addressed. Our approach focuses on regulating the depth of hypnosis and hemodynamic stability, controlling propofol and remifentanil. However, a comprehensive anesthesia controller should also manage additional factors such as NMB and vasopressors. In addition, the effectiveness of our agent has not been tested in a clinical setting. The practical applicability of this study is inherently tied to the simulator's accuracy, which, despite its complexity, cannot fully capture the variability of human physiology. Furthermore, no safety constraints were incorporated into the agent's decision-making, which is critical for future deployment in a real-world setting. Upcoming research could focus on developing more accurate physiological models, collecting real-world data, integrating safety mechanisms, and expanding the control objectives to other aspects of anesthesia management to ensure this approach can be effectively and safely used in clinical practice.

References

[1] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35 th International Conference on Machine Learning*, volume 80, pages 1587–1596, Stockholm, Sweden, October 2018.

[2] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, November 2017.

[3] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

[4] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley-Interscience, 2014.

[5] Richard S. Sutton and Andrew Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts London, England, second edition edition, 2020.