*mathematics*

MDPI

*Article*

# Trust Region Policy Learning for Adaptive Drug Infusion with Communication Networks in Hypertensive Patients

**Mai The Vu [1], Seong Han Kim [1], Ha Le Nhu Ngoc Thanh [2], Majid Roohi [3] and Tuan Hai Nguyen [4,*]**

[1] Department of Artificial Intelligence and Robotics, Sejong University, Seoul 05006, Republic of Korea; maithevu90@sejong.ac.kr (M.T.V.); shkim8@sejong.ac.kr (S.H.K.)
[2] Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology and Education, Ho Chi Minh City 71307, Vietnam; thanh.hlnn@hcmute.edu.vn
[3] Department of Mathematics, Aarhus University, 8000 Aarhus, Denmark; mr@math.au.dk
[4] Faculty of Engineering, Dong Nai Technology University, Bien Hoa City, Vietnam
* Correspondence: nguyentuanhai@dntu.edu.vn

**Abstract:** In the field of biomedical engineering, the issue of drug delivery constitutes a multifaceted and demanding endeavor for healthcare professionals. The intravenous administration of pharmacological agents to patients and the normalization of average arterial blood pressure (AABP) to desired thresholds represents a prevalent approach employed within clinical settings. The automated closed-loop infusion of vasoactive drugs for the purpose of modulating blood pressure (BP) in patients suffering from acute hypertension has been the focus of rigorous investigation in recent years. In previous works where model-based and fuzzy controllers are used to control AABP, model-based controllers rely on the precise mathematical model, while fuzzy controllers entail complexity due to rule sets. To overcome these challenges, this paper presents an adaptive closed-loop drug delivery system to control AABP by adjusting the infusion rate, as well as a communication time delay (CTD) for analyzing the wireless connectivity and interruption in transferring feedback data as a new insight. Firstly, a nonlinear backstepping controller (NBC) is developed to control AABP by continuously adjusting vasoactive drugs using real-time feedback. Secondly, a model-free deep reinforcement learning (MF-DRL) algorithm is integrated into the NBC to adjust dynamically the coefficients of the controller. Besides the various analyses such as normal condition (without CTD strategy), stability, and hybrid noise, a CTD analysis is implemented to illustrate the functionality of the system in a wireless manner and interruption in real-time feedback data.

**Keywords:** average arterial blood pressure (AABP); vasoactive drugs; nonlinear backstepping controller (NBC); model-free reinforcement learning (MF-DRL); communication time delay (CTD)

**MSC:** 37N25; 37N35; 93A99; 93C10; 93C15; 93C83; 93C95

check for updates

Academic Editors: Óscar Valero Sierra and Yong Xie

Received: 27 October 2024
Revised: 17 December 2024
Accepted: 29 December 2024
Published: 1 January 2025

**Citation:** Vu, M.T.; Kim, S.H.; Thanh, H.L.N.N.; Roohi, M.; Nguyen, T.H. Trust Region Policy Learning for Adaptive Drug Infusion with Communication Networks in Hypertensive Patients. *Mathematics* **2025**, *13*, 136. https://doi.org/10.3390/math13010136

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Patients undergoing major surgery (especially cardiovascular surgery) may experience episodes of acute blood pressure elevation during the intra- and postoperative periods. Such episodes may last between a few minutes and several hours and involve a heightened risk of complications for the patient, such as cardiovascular or cerebrovascular accidents. Careful management of acute hypertension is therefore required and is generally achieved by infusion of suitable vasoactive drugs by clinical operators [1–4].

Due to significant variability in pharmacological sensitivity among patients and even within the same individual over temporal spans, the appropriate infusion rate cannot

be established a priori, necessitating the process of dose titration. This undertaking is labor-intensive and time-consuming for clinical practitioners, and manual titration may lack precision. In this framework, the advancement of a closed-loop system capable of autonomously modulating the drug infusion rate to sustain a patient's average arterial blood pressure (AABP) within a designated range would constitute a noteworthy enhancement to medical technology [5,6]. The advantages of such a system would encompass an elevation in patient outcomes by diminishing the probability of complications, eliminating the risk of human error, and enabling the reallocation of human resources towards more critical responsibilities. This study addresses the challenge associated with the automatic closed-loop regulation of sodium nitroprusside (SNP) infusion. SNP is a potent and rapidly acting vasodilatory agent that facilitates a reduction in blood pressure by inducing a decrease in both arterial and venous resistance. Patients exhibiting elevated blood pressure subsequent to surgical procedures have conventionally received treatment with this particular medication [7–9].

The complexities associated with the response of the SNP medication paradigm, encompassing time-dependent uncertainties, modeling discrepancies, and temporal delays, render the automation of the control of a patient's AABP a formidable challenge. In prior scholarly investigations, researchers have devised an array of systems intended for the modulation of AABP through the utilization of SNP. For example, earlier studies [10,11] presented an optimal Interval Type-2 Fuzzy Logic Control (IT2-FLC) framework alongside a two-layer Fractional Order Fuzzy Logic Control, wherein the optimization of controller parameters was conducted via the Grey Wolf Optimization (GWO) algorithm. Reference [12] executed an IT2-FLC system with the optimization of controller parameters being realized through the Cuckoo Search algorithm. Furthermore, in an alternative investigation, a Model Predictive Control (MPC) strategy employed Particle Swarm Optimization (PSO) for the calibration of the controller's coefficients [13]. Additionally, a Fractional Order Proportional Integrative (FO-PI) controller was conceived to govern MABP. In another distinct study, the Q-learning algorithm, a variant of reinforcement learning, and the active disturbance rejection control (ADRC) utilizing deep deterministic policy gradient (DDPG) for the tuning of ADRC coefficients were implemented as a modern control approach for MABP regulation [14,15]. In addition to these methodologies, other techniques such as the development and assessment of a simple adaptive PI controller for MAP regulation [16], robust multi-model adaptive control [17], and embedded PI controller [18] have been devised for the management of BP.

In prior research, significant advancements have been made in regulating AABP using various control methodologies; however, several notable deficiencies persist. Firstly, most studies rely on ideal connection assumptions, neglecting the complexities of real-world implementations that often depend on wireless connectivity, which can introduce signal delays, and communication variability, adversely affecting control efficacy. Secondly, many controllers, such as MPC and PID, depend heavily on mathematical models that may not accurately represent the system dynamics, resulting in performance declines when actual conditions deviate from these models. Thirdly, while IT2-FLC does not require exact models, it faces challenges related to rule complexity, tuning difficulties, and interpretability [19,20]. Additionally, metaheuristic optimization algorithms like GWO and PSO can struggle with local optima convergence and parameter sensitivity, while DRL methods, such as Q-learning, suffer from sample inefficiency and exploration–exploitation trade-offs. Although the implementation of these strategies may result in favorable outcomes for a finite period, it is imperative to acknowledge that they do not ensure either optimal or locally optimal solutions. The related constraints encompass limited generalization abilities, lack of learning ability, convergence to local optima, and an increased susceptibility to vari-

ations in parameter configurations [21,22]. To address these issues, this study introduces a novel data-driven NBC optimized using TRPO. This approach reduces reliance on precise mathematical models, incorporates communication time delays to enhance robustness in wireless scenarios, and effectively manages model uncertainties. The TRPO algorithm improves parameter tuning, ensuring better exploration–exploitation balance and real-time adaptability, thereby enhancing the controller's performance in dynamic environments where system parameters may vary. The structure of the closed-loop AABP strategy is presented in Figure 1.

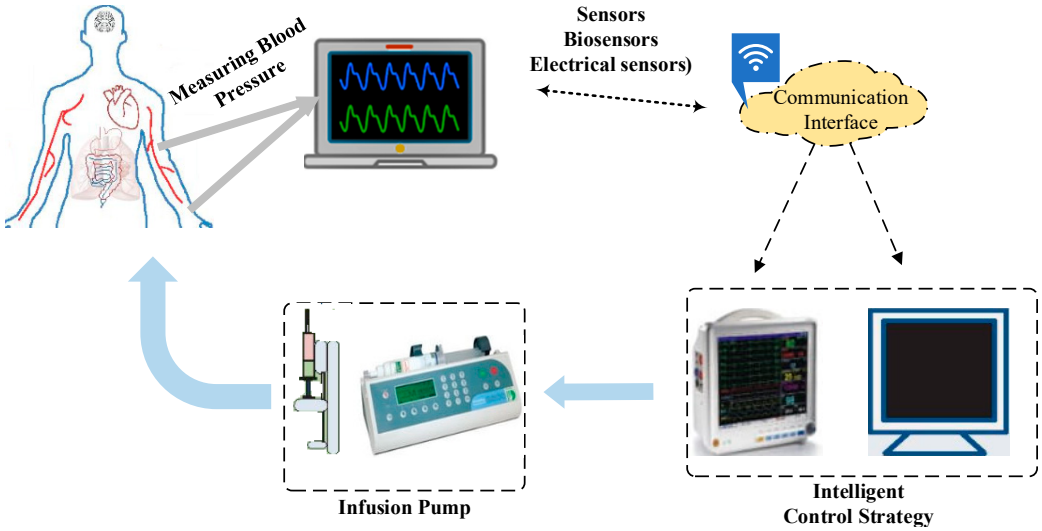



**Figure 1.** The overall structure of the closed-loop AABP mechanism.

The backstepping controller has conventionally been conceived as a model-based control strategy, distinguished by its methodical design process that utilizes mathematical models of the system to guarantee stability and performance. This methodology entails the iterative formulation of control laws predicated on the system's dynamics, facilitating accurate tracking and regulation of intended outputs. Nonetheless, in this manuscript, we present an innovative model-free backstepping controller that considerably augments flexibility and adaptability. In contrast to its model-based analogue, this model-free variant is not contingent upon precise mathematical characterizations of the system, thereby empowering it to proficiently manage uncertainties and nonlinearities in real-time applications. Through the application of a data-driven approach, the model-free backstepping controller is capable of continuously assimilating insights from the system's behavior, enabling real-time modifications and enhanced performance across diverse operating conditions. This progression not only alleviates the constraints associated with model inaccuracies but also bolsters the controller's resilience in dynamic environments.

The complex attributes of physiological dynamic systems present a considerable challenge in the formulation of controllers for previously unexamined systems. Additionally, fluctuations in the output may elicit significant alterations in the system's behavior. The nonlinear backstepping control (NBC) methodology guarantees the global asymptotic stability of stringent feedback systems by creating a linkage between the selection of a control Lyapunov function and the development of a feedback controller [23,24]. By integrating parameter estimations for unidentified variables within the control framework, the adaptive NBC approach emerges as a refined iteration of the backstepping control paradigm, appropriate for a variety of applications. In contexts characterized by uncertainties, the NBC strategy is recognized as an invaluable instrument. Notably, NBC has exhibited exceptional effectiveness in governing nonlinear systems with fixed models in the presence of external disturbances [25]. Furthermore, the implementation of integral action functions to mitigate

steady-state discrepancies is significant. It is crucial to underscore that NBC operates in a recursive fashion [26]. Despite the numerous advantages offered by the NBC relative to both Conventional Controllers and modern techniques, it is pertinent to acknowledge that this specific controller is constrained by the lack of online functionality and automatic parameter adjustment. This limitation represents a considerable barrier when employing the NBC in complex dynamic systems and biological simulations. The absence of online adaptability and automated parameter modifications may hinder the controller's capacity to adeptly navigate the inherent complexities, practical implementation issues, and time-sensitive characteristics of such systems. To enable online updates for the developed strategy, a model-free deep reinforcement learning (MF-DRL) framework, specifically the Trust Region Policy Optimization (TRPO) algorithm, has been designed and incorporated into the NBC as an on-policy approach [27].

TRPO represents a resilient algorithm within the domain of reinforcement learning (RL) that has attracted significant scholarly attention within the field of artificial intelligence (AI) [28]. A pivotal aspect of TRPO lies in its ability to provide consistent and incremental improvements to the policy during the optimization process [29]. Unlike other policy-based methodologies that may face instability and oscillations, TRPO incorporates a trust region constraint to ensure that the newly modified policy remains sufficiently close to its predecessor, thus preventing drastic declines in performance. This feature makes TRPO particularly advantageous for tasks that involve large state and action spaces, where navigating the policy landscape can be particularly challenging. The core principle supporting TRPO centers on optimizing the policy by maximizing the expected return within a defined trust region, with this limitation characterized by the Kullback–Leibler (KL) divergence between the existing and updated policies. This approach enables TRPO to identify the most effective policy adjustment while ensuring that changes in the policy distribution remain within a manageable range, thereby promoting stable learning outcomes and enhanced sample efficiency [30,31]. In the current investigation, an advanced methodology utilizing the NBC in conjunction with the TRPO algorithm has been formulated to rectify the shortcomings identified in previous research endeavors. Furthermore, a novel perspective has been introduced by incorporating a communication time delay into the established framework, intended to illustrate the operational capabilities of closed-loop AABP within a wireless context. The contributions of the present manuscript are encapsulated as follows:

(1) The formulation of a data-driven nonlinear backstepping controller (NBC) is intended to concurrently modulate the average arterial blood pressure (AABP) via vasoactive drugs, thereby facilitating a more precise and efficient regulation of the system in comparison to traditional controllers that rely on static, predetermined configurations.
(2) The implementation of the MF-DRL methodology serves to augment the effectiveness of the NBC strategy through the adjustment of parameters. This technique enhances the controller's performance and permits timely modifications, thereby ensuring the system's capacity to adaptively respond to fluctuating conditions and continuously improve its operational capabilities.
(3) The implementation of the communication time delay (CTD) approach allows for an in-depth analysis of how transfer delay signals impact the system's dynamics. This framework facilitates a comprehensive evaluation of its effectiveness and efficiency in managing and mitigating these delays, which is crucial for real-world applications. As a result, it offers valuable insights into the system's robustness and applicability across various latency scenarios.
(4) The performance of the established system when faced with a variety of patients is analyzed, in addition to assessing the system's robustness against variations in critical parameters and exposure to hybrid noise disturbances.

The organization of this paper is as follows. The problem of average arterial blood pressure (AABP) is mathematically formulated in Section 2. Moreover, the control methodology is presented in this section. The next section, Section 3, is devoted to the design of the controller using the TRPO algorithm. The simulation examinations are provided in Section 4. The results of the proposed methodology are concluded in Section 5.

## 2. Problem Formulation and Methodology

### *2.1. AABP Dynamic Model*

This section delineates the dynamic patient model utilized for the continuous monitoring of average arterial blood pressure (AABP) during the administration of the sodium nitroprusside (SNP) pharmacological agent, aimed at achieving the patients' targeted AABP. Furthermore, the ensuing discourse encompasses the limitations inherent in the control objectives of the previously mentioned model. The dynamic model under examination elucidates the relationship between the infusion rates of the SNP therapeutic agent and the subsequent alterations in mean AABP as a consequence of the drug administration. The dynamic model elucidates how fluctuations in BP influence the velocity at which the pharmacological agent is administered. The representation of the dynamic model is articulated as follows [16,32]:

$$\mathcal{G}_{patient}(\mathfrak{z}) = \frac{\mathcal{Y}_{patient}(\mathfrak{z})}{\mathcal{I}_{patient}(\mathfrak{z})} = \left( \frac{\psi_{patient}(1 + X_{i3}\mathfrak{z})e^{-(\Omega_{patient})\mathfrak{z}}}{\left[(1 + X_{i3}\mathfrak{z})(1 + X_{i2}\mathfrak{z}) - \varpi_{patient}\right]} \right) \times \left( \frac{1}{1 + X_{i1}\mathfrak{z}} \right) \quad (1)$$

The patient's blood pressure change, represented by $\mathcal{Y}_{patient}(\mathfrak{z})$, is the controlled variable, while the rate of drug infusion, $\mathcal{I}_{patient}(\mathfrak{z})$, is the control input. The temporal constants $X_{i1}\mathfrak{z}$, $X_{i2}\mathfrak{z}$, and $X_{i3}\mathfrak{z}$ dictate the rate at which the SNP medication is assimilated and takes effect in the patient's physiological system for different individuals. The parameter $\varpi_{patient}$ represents the proportion of recirculated SNP medication that contributes to the overall therapeutic effect, while $\Omega_{patient}$ captures the time delay between the initiation of medication administration and its impact on the patient's blood pressure. The parameter $\psi_{patient}$ illustrates the patient's sensitivity to the pharmacological intervention and its subsequent effect on their blood pressure. The closed-loop structure allows the system to continuously monitor the patient's blood pressure and adjust the drug infusion rate accordingly to maintain the desired AABP level, with the nonlinear transfer functions capturing the complex dynamics involved in the drug administration and its impact on the patient's physiology. Consequently, the output of the AABP can be articulated as follows:

$$AABP(t) = S_s + AABP(0) \quad (2)$$

The initial BP denoted as $AABP(0)$. The comprehensive structure of the studied model is shown visually in Figure 2a.

As previously articulated, the primary objective of this study is to regulate AABP through meticulous administration of the medication SNP. Within the framework of this research, the reference input value, as depicted in Figure 2b, necessitates effective control via the appropriate dosage of SNP medication to achieve an AABP that remains within the acceptable threshold of 100 mmHg. Given the deviation from the initial elevated AABP value of 150 mmHg, the constraints associated with the SNP drug infusion rate must be judiciously considered.

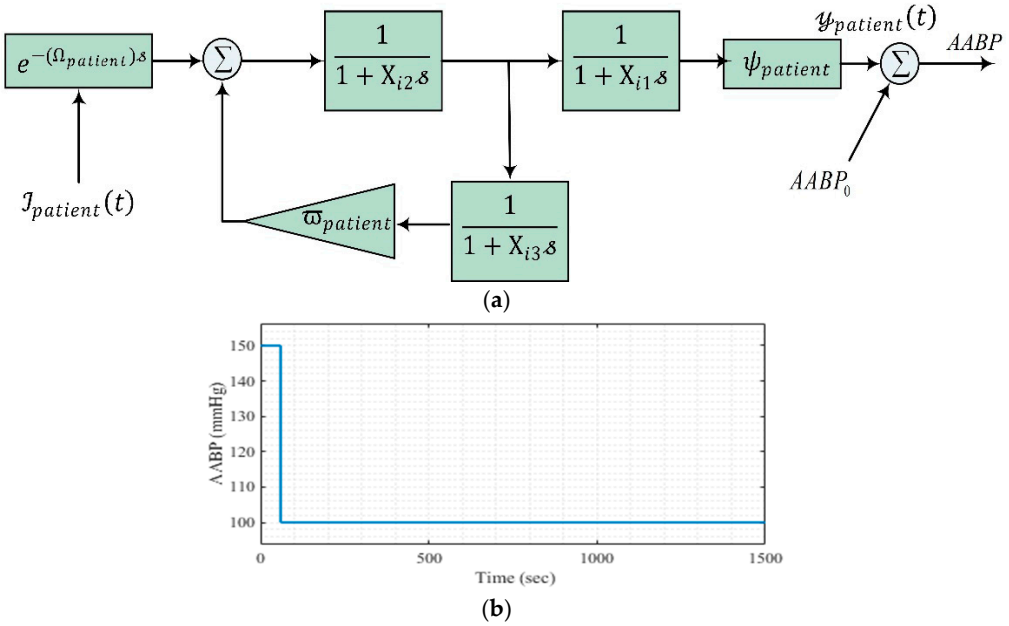



**Figure 2.** Dynamic model of the AABP system with NBC and the reference input: (**a**) Dynamic model, (**b**) reference input.

### *2.2. Mechanism of Nonlinear Backstepping Controller (NBC)*

The NBC's structure is depicted in Figure 2a, and the controller's specifics are provided below. The NBC approach and system stability will be discussed next. The following model is used to demonstrate the BP dynamics:

$$n^{(x)} = \beta(n) + \kappa u \tag{3}$$

where $\beta(n)$ and $u$, respectively, indicate the BP and input system's dynamics.

$$n^{(x)} = \beta(n) + \beta_e(\cdot) + \hat{\kappa} u \tag{4}$$

Where $\hat{\kappa}$ is the estimate of the unknown gain of the parameter $M$, and $\beta_e(\cdot)$ is the unmodeled and uncertain dynamics of BP. Thus, $\beta_e(\cdot)$ can be written as follows: $\beta_e(\cdot) = \Lambda + (\kappa - \hat{\kappa})u$. The model uncertainties are represented by $\Lambda$. An observer can be used for the known and projected nonlinear dynamics of BP, which can reduce the error of some state variables [33].

$$n^{(x)} = \beta(n) + F + \hat{\kappa} u \tag{5}$$

$$F = n^{(x)} - f(n) - \hat{\kappa} u \tag{6}$$

$$u = \frac{F - n_d^{(x)} - u_c}{\hat{\kappa}} \tag{7}$$

$$n^{(x)} = \beta(n) + n_d^{(x)} + u_c \tag{8}$$

The AABP model's state variable can be represented as follows:

$$n_1 = AABP \tag{9}$$

$$n_2 = \dot{n}_1 \tag{10}$$

Nevertheless, in fact, the desired and actual values of the state variable ($n_1$) differ; this difference is expressed by the formula $E_1 = n_1 - n_1^d = n_1 - n_d$. The NBC theory can

be used to force position-tracking and amplitude-tracking error to converge to a single variable. The Lyapunov function is selected in order to ensure the convergence stability of the nonlinear BP model. The Lyapunov function $L(\mathrm{M}_1)$ can be expressed as follows and will be positive definite around the state variable:

$$L_1(\mathrm{M}_1) = \frac{1}{2}\mathrm{M}_1^2 \tag{11}$$

The derivative of Equation (11) can be written as follows:

$$\dot{L}_1(\mathrm{M}_1) = \mathrm{M}_1\dot{\mathrm{M}}_1 = \mathrm{M}_1\left(n_2 - \dot{n}_d\right) \tag{12}$$

A dynamic disparity occurs between a variable and its intended value because we are not able to regulate the variable. As a result, the amplitude-tracking error can be used to offset the dynamic error.

$$\mathrm{M}_2 = n_2 - n_2^d \tag{13}$$

In the event that a semi-negative Lyapunov function is selected, the error will approach zero. The following is one way to express the implicit input:

$$n_2^d = \dot{\mathrm{n}}_d - w_1\mathrm{M}_1 \tag{14}$$

$$\mathrm{M}_2 = n_2 - \dot{\mathrm{n}}_d + w_1E_1 \tag{15}$$

Errors and uncertainties in the model lead to steady-state inaccuracy. This problem can be resolved by applying the integral term to the system as described below [34].

$$n_2 = \dot{\mathrm{n}}_d - w_1\mathrm{M}_1 - w_3\int \mathrm{M}_1 \tag{16}$$

$$\mathrm{M}_2 = n_2 - \dot{\mathrm{n}}_d + w_1\mathrm{M}_1 + w_3\int \mathrm{M}_1 \tag{17}$$

Therefore, the following information describes the derivative for both amplitude and position tracking.

$$\dot{\mathrm{M}}_1 = \dot{\mathrm{n}}_1 - \dot{\mathrm{n}}_d = \mathrm{M}_2 + n^* - \dot{\mathrm{n}}_d = \mathrm{M}_2 - w_1\mathrm{M}_1 - w_3\int \mathrm{M}_1 \tag{18}$$

$$\dot{\mathrm{M}}_2 = \dot{n}_2 - \ddot{n}_d + w_1\dot{\mathrm{M}}_1 - w_3e\mathrm{M}_1 = \dot{n}_2 - \ddot{n}_d + w_1\mathrm{M}_2 - w_1^2\mathrm{M}_1 - w_1w_3\int E_1 - w_3\mathrm{M}_1 \tag{19}$$

The Lyapunov functions $L_1(\mathrm{M}_1.\int \mathrm{M}_1)$ and $L_2(\mathrm{M}_1.\mathrm{M}_2.\int \mathrm{M}_1)$ are defined for the position- and amplitude-tracking error and are expressed as follows [35]:

$$L_1\left(\mathrm{M}_1.\int \mathrm{M}_1\right) = \frac{1}{2}\mathrm{M}_1^2 + \frac{w_3}{2}\left(\int \mathrm{M}_1\right)^2 \tag{20}$$

$$\dot{L}_1\left(\mathrm{M}_1.\int \mathrm{M}_1\right) = \mathrm{M}_1\dot{\mathrm{M}}_1 + w_3\mathrm{M}_1\int \mathrm{M}_1 = \mathrm{M}_1\mathrm{M}_2 - w_1{\mathrm{M}_1}^2 \tag{21}$$

$$L_2\left(\mathrm{M}_1.\mathrm{M}_2.\int \mathrm{M}_1\right) = \frac{1}{2}\mathrm{M}_1^2 + \frac{1}{2}\mathrm{M}_2^2 + \frac{w_3}{2}\left(\int \mathrm{M}_1\right)^2 \tag{22}$$

$$\begin{aligned}\dot{L}_2(\mathrm{M}_1.\mathrm{M}_2.\int \mathrm{M}_1) &= \mathrm{M}_1\dot{\mathrm{M}}_1 + \mathrm{M}_2\dot{\mathrm{M}}_2 + w_3\mathrm{M}_1\int \mathrm{M}_1 \\ &= \mathrm{M}_1\mathrm{M}_2 - w_1{\mathrm{M}_1}^2 + \mathrm{M}_2\left(\dot{n}_2 - \ddot{n}_d + w_1\mathrm{M}_2 - w_1^2\mathrm{M}_1 - w_1w_3\int \mathrm{M}_1 - w_3\mathrm{M}_1\right)\end{aligned} \tag{23}$$

In order to ensure that the $E_2$ will eventually approach zero, the $\dot{L}_2(\mathrm{M}_1.\mathrm{M}_2.\int \mathrm{M}_1)$ must be semi-negative definite. The solution to this is to select Equation (24).

$$\mathrm{M}_1\mathrm{M}_2 + \mathrm{M}_2\left(\dot{n}_2 - \ddot{n}_d + w_1\mathrm{M}_2 - w_1^2\mathrm{M}_1 - w_1w_3\int \mathrm{M}_1 - w_3\mathrm{M}_1\right) = -w_2{\mathrm{M}_2}^2 \tag{24}$$

Thus, $\dot{n}_2$ can be expressed as follows:

$$\dot{n}_2 = f(x) + \ddot{n}_d + u_c = \ddot{n}_d + \left(w_1^2 + w_3 - 1\right)\mathrm{M}_1 - (w_1 + w_2)\mathrm{M}_2 + w_1w_3\int \mathrm{M}_1 \tag{25}$$

$$u = \left[\left(w_1^2 + w_3 - 1\right)\mathrm{M}_1 - (w_1 + w_2)\mathrm{M}_2 + w_1w_3\int \mathrm{M}_1\right] - \beta(x) \tag{26}$$

**Remark 1.** *The performance of the backstepping controller relies heavily on the careful selection and precise tuning of its coefficients. This tuning is crucial for enhancing the controller's effectiveness and stability, allowing it to optimally respond to varying system dynamics. Therefore, thorough adjustment of these coefficients is essential for achieving desired control objectives and ensuring robustness in practical applications.*

**Remark 2.** *Traditional methods for tuning NBC coefficients, such as heuristic and fuzzy approaches, have notable weaknesses. Heuristic methods often rely on trial and error, leading to time-consuming processes without guaranteed optimal solutions. Fuzzy Logic tuning can become complex and difficult to manage, especially with multiple input variables, resulting in challenges related to rule complexity and interpretability. To address these issues, we have developed the TRPO method, which offers a systematic and data-driven approach for tuning NBC coefficients. TRPO enhances the tuning process by ensuring stable learning and convergence to optimal solutions, ultimately improving the performance and robustness of the backstepping controller.*

## 3. The Design of the NBC Based on Trust Region Policy Optimization DRL

The Trust Region Policy Optimization (TRPO) algorithm is a model-free strategy that represents a prominent technique within the domains of artificial intelligence (AI) and machine learning (ML). This algorithm is meticulously crafted to augment the policy of an agent by incrementally improving its operational efficacy in a stable and effective fashion [36]. A fundamental characteristic of TRPO is its focus on ensuring that only minor adjustments to the policy occur during each iteration, which functions to maintain stability and avert significant modifications that could detrimentally impact the learning trajectory [37]. This technique aspires to derive an efficient behavioral policy through iterative trial-and-error interactions with an opaque environment. The primary aim is to optimize an agent's behavioral policy concerning the overarching expected discounted reward. MF-DRL algorithms do not formulate a model of the environment. This algorithm is generally adaptable, requires relatively minimal modifications, and is capable of seamlessly integrating powerful function approximators such as deep neural networks (DNNs). The topology of the policy network and value network of the TRPO algorithm is illustrated in Figure 3.

TRPO is categorized as an on-policy algorithm. On-policy approaches confer a multitude of advantages, including the ability to evaluate each resultant policy [38,39]. Furthermore, in the context of on-policy optimization, traditional policy gradient algorithms experience sporadic updates characterized by substantial step sizes, resulting in the acquisition of inferior samples. To mitigate this instability, this methodology constrains the Kullback–Leibler (KL) divergence between successive policies to facilitate significantly more stable updates [40].

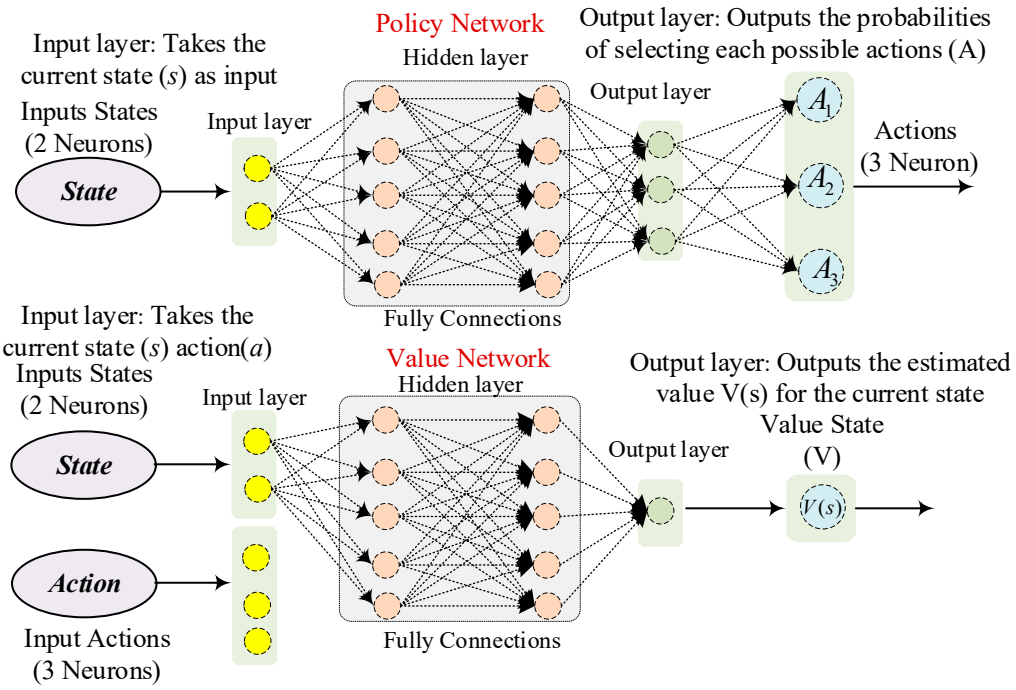



**Figure 3.** The structure of the policy and value networks.

### *3.1. The Fundamental of TRPO Algorithm*

The expected return from the start state is represented by the policy performance objective $\xi(\pi) = E_{s0,\, a0}\,[Rew_0]$, and the goal of RL is to maximize it. This objective is stated as follows:

$$\xi(\pi) = E_{s0,\, a0,\ldots}\left[\sum_{0}^{\infty} \lambda^t r(X_t,\, a_{kt})\right] \tag{27}$$

$$X_0 \sim \delta_0 \sim a_t \sim \pi(0,\, X_t),\; X_{t+1} \sim \mathrm{Y}(0, X_t,\, a_t) \tag{28}$$

where the state and action spaces are denoted, respectively, by $X$ and $a$. Furthermore, $r$ denotes the reward term, $\delta$ represents the initial state distribution, Y represents the dynamic distribution, and $\lambda$ determines the discount factor [41]. The advantage over $\pi$, accumulated over time, can be utilized to predict the expected return on investment for an additional policy $\hat{\pi}$:

$$\xi(\hat{\pi}) = \xi(\pi) + E_{s0,\, a0,\ldots\sim\hat{\pi}}\left[\sum_{0}^{\infty} \lambda^t A_\pi(X_t,\, a_{kt})\right] \tag{29}$$

$$= \xi(\pi) + \sum_X \delta_{\hat{\pi}}\,(X) \sum_a \hat{\pi}(a|X) A_\pi(X, a) \tag{30}$$

It is crucial to recognize that any policy update $\pi \to \hat{\pi}$ that satisfies Equation (31) guarantees the improvement of policy $\xi$. The guarantee of policy improvement is difficult because $\hat{\pi}$ influences $\delta_{\hat{\pi}}(X)$. This leads to complexity. To reduce such interdependence, a possible solution is provided by a local estimation of $\delta_{\hat{\pi}}$, where $\delta_\pi$ is substituted for the visiting frequency $\xi(\hat{\pi})$ [42].

$$\sum_X \delta_{\hat{\pi}}\,(X) \sum_a \hat{\pi}(a|X) A_\pi(X, a) \geq 0 \tag{31}$$

$$\mathrm{T}_\pi(\hat{\pi}) = \xi(\pi) + \sum_X \delta_\pi\,(X) \sum_a \hat{\pi}(a|X) A_\pi(X, a) \tag{32}$$

Improvements to $\mathrm{T}_\pi$ will result in improvements to $\xi$ if the perturbation $\pi \to \hat{\pi}$ is negligibly large. However, this small correction places a limit on the possible growth of $\xi$.

The TRPO algorithm is suggested as a workaround for this restriction, driven by the ideas of conservative policy iteration [43]. TRPO introduces a surrogate objective function and shows how improving this function improves policy performance by placing a more relaxed constraint on policy performance. In particular, TRPO provides the following lower bound:

$$\xi(\pi_{new}) \geq \mathrm{T}_{\pi_{old}}(\pi_{new}) - \frac{4 \in \lambda}{(1-\lambda)^2} d_{KL}^{max}(\pi_{old},\ \pi_{new}) \tag{33}$$

where $\in = max_{X,a}\left|A_{\pi_{old}}(X,\ a)\right|$. The existing policies, identified as $\pi_{old}$ and $\pi_{new}$, are separated by a measure called $d_{KL}^{max}(\pi_{old},\ \pi_{new})$, which is determined by the KL divergence:

$$d_{KL}^{max}(\pi_{old},\ \pi_{new}) = max_X d_{KL}(\pi_{old}(0|X)||\pi_{new}(0|X)) \tag{34}$$

As shown by $\xi$, Equation (34) shows that improving the surrogate objective function on the right-hand side may enhance policy performance. The surrogate objective function in $\mathrm{T}_{\pi_{old}}(\pi_{new})$ is dependent on a particular behavior policy, namely, the present policy $\pi_{\mathrm{old}}$, for the frequency of visits $\delta_{\pi_{old}}$. The process of the functionality of the TRPO strategy is illustrated in Figure 4.

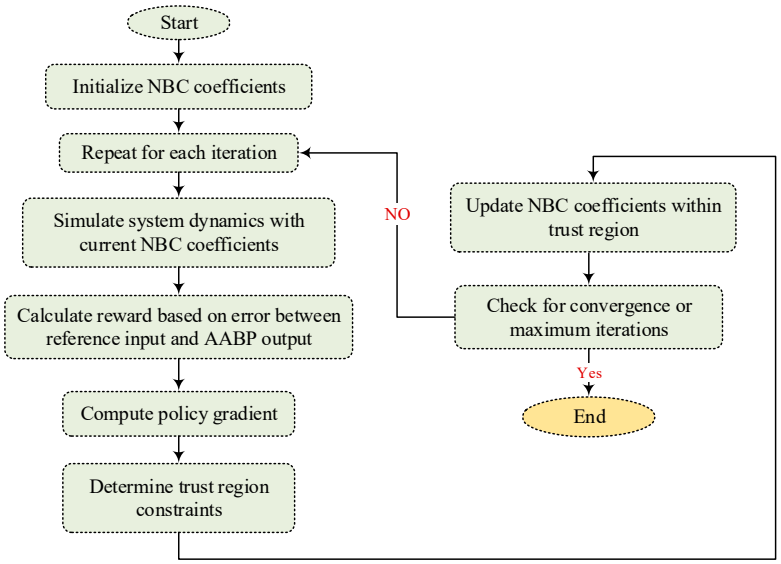



**Figure 4.** The flowchart of the TRPO algorithm.

### *3.2. The Functionality of the TRPO Algorithm in Tuning NBC's Coefficients*

The efficacy of the policy methodology can be deduced from the magnitude of the reward associated with the agent's state–action pair. As elucidated in Section 2.1, a relationship exists between the values of both variables (AABP and SNP); hence, both variables can be incorporated into the reward function. For this purpose, as illustrated in Figure 5, the error between the reference input and AABP output is considered as a reward function for the specifically designed controller. Consequently, the final objective is to ascertain that the algorithm endeavors to modulate the NBC coefficients by actions ($A_1$, $A_2$, $A_3$) and to manage the level of AABP produced. The prevailing policy (i.e., the existing configuration of controller parameters) is systematically assessed by simulating the system dynamics and quantifying the reward, subsequently enhancing the policy utilizing the TRPO algorithm. This methodology is applied to each of the three NBC parameters. To preserve stability, the TRPO algorithm computes the policy gradient, which elucidates the optimal direction for updating the policy parameters to optimize reward. It subsequently imposes constraints on the updated step size to ensure that the new policy remains sufficiently aligned with the preceding version. Subsequently, while adhering to the trust region constraints, the policy

parameters are modified in accordance with the predicted policy gradient. Through the iterative execution of this protocol, the TRPO algorithm is capable of identifying, from the simulated dynamic model, the optimal controller parameters that manage AABP while concurrently considering SNP dose control.

$$\mathbb{R} = \frac{1}{(Ref - AABP)^2} \tag{35}$$

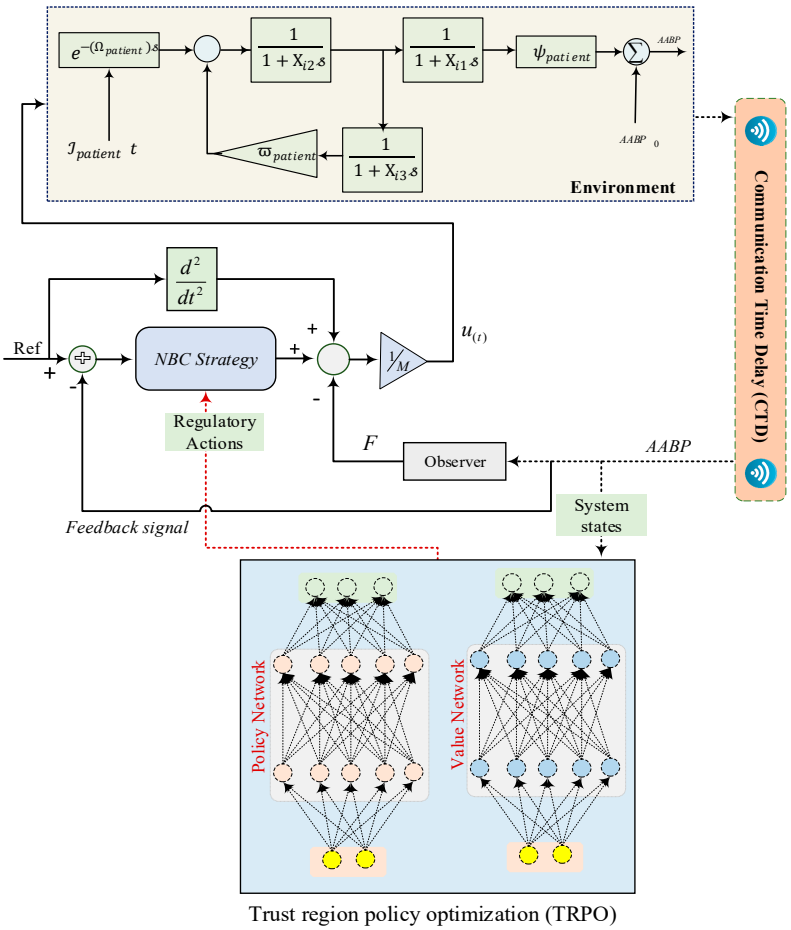



**Figure 5.** The overall structure of tuning NBC's coefficients by TRPO with CTD technique.

From Figure 5, it can be seen that the fundamental component of the proposed framework is the NBC strategy, which is tasked with the generation of regulatory actions that are subsequently administered to the patient. This NBC assimilates the reference input (Ref) alongside the current states of the system, subsequently yielding the requisite control signal to be implemented for the patient. Upon the generation of the control signal by the NBC, it is subsequently administered to the patient, precipitating alterations in the patient's dynamics and the comprehensive state of the system. The modified state of the system is then monitored by the observer block, which accumulates pertinent information regarding the system, notably the AABP signal emanating from the environment. The acquired state information of the observed system is then reintroduced into the TRPO component. The TRPO algorithm employs this state information to refine both the policy network and the value network, which constitute the principal elements of the TRPO algorithm. The policy network is tasked with the mapping of system states to the optimal actions, specifically the regulatory actions produced by the NBC. Conversely, the value network is designed to estimate the anticipated long-term reward (or value) associated with being in a given state while executing a specific action. The TRPO algorithm subsequently utilizes these policy

and value networks to engage in policy optimization. This is accomplished by modifying the weights and biases of the networks through techniques based on gradient optimization, while concurrently ensuring that the policy updates remain within acceptable limits relative to the existing policy, adhering to the "trust region" constraint. Through the iterative refinement of the policy and value networks, the TRPO algorithm seeks to ascertain the optimal policy capable of effectively managing the patient's condition, considering the intricate dynamics involving the patient, their environment, and the interactions therein.

## 4. Simulation Verifications and Statistical Analysis

This section provides a detailed examination of the efficacy of the TRPO-based NBC methodology within the dynamic framework of the AABP test system, aiming to handle BP by adjusting SNP dosage under normal conditions (without CTD) as well as in wireless interactions among individuals diagnosed with high BP. The nominal value of the AABP system and hyperparameters of the TRPO algorithm have been adjusted according to Table 1.

**Table 1.** Quantities and variables of the AABP system and TRPO hyperparameters.

| Variables | Quantities | Variables | Quantities |
|---|---|---|---|
| Systematic circulation | $X_{i2} = 30$s | Drug fraction recirculated | $\Phi = 0.5$ |
| Drug action | $X_{i1} = 50$s | Delay | $(\Omega)$s $= 60$s |
| Pulmonary Circulation | $X_{i3} = 10$s | Sensitivity | $P = -0.25$ mmHg |
| **TRPO Hyperparameters** | | | |
| Number of hidden layers | 2 | Activation function | ReLU |
| Learning rate | 0.01 | Discount factor | 0.99 |
| KL divergence constraint | | 0.01 | |

To achieve the most advantageous outcome, it is imperative to acknowledge the limitations imposed by the administration of SNP medications [12,14]. The subsequent constraints are enumerated:

**Constraint 1.** *The definitive closed-loop control mechanism must not exhibit oscillatory behavior at any point in time (oscillations within the system governing (AABP) possess the potential to induce instability and adverse consequences for the patient. The preservation of stability is imperative to ensure an uninterrupted and secure regulation of AABP).*

**Constraint 2.** *The proposed sensor should achieve stabilization within a time frame of less than ten to fifteen minutes, respectively (the attainment of rapid stabilization of the AABP control system is vital for promptly reaching the designated AABP target value and minimizing the duration of suboptimal AABP encountered during the transient response phase).*

**Constraint 3.** *The optimal AABP should be maintained within the range of [70, 120] mmHg subsequent to stabilization (the AABP range of 70 to 120 mmHg is universally recognized in clinical practice as an appropriate target for cardiac patients in the postoperative recovery phase, considering its congruence with safety parameters and physiological standards that facilitate patient rehabilitation).*

**Constraint 4.** *It is imperative to ensure that the patient's AABP does not fall below the critical threshold of 70 mmHg (maintaining AABP levels above 70 mmHg is crucial to ensure adequate perfusion and oxygenation to vital organs, thereby preventing the emergence of potential complications such as organ dysfunction or insufficient perfusion* [44]).

**Constraint 5.** *In order to attain the desired AABP of 100 mmHg, it is critical that the ultimate steady-state MABP remains confined within a tolerance of ±5 mmHg (sustaining the AABP within a range of ±5 mmHg from the reference value of 100 mmHg is essential for ensuring precise regulation and mitigating deviations that may adversely affect the patient's health status).*

**Constraint 6.** *Administering a substantial quantity of sodium nitroprusside (SNP) to a patient undergoing treatment or recuperation from an injury may pose significant risks. Consequently, it is vital that the patient's medication dosage be constrained between 0 and 180 milliliters per hour. The methodology, along with the testing system, has been systematically implemented into Simulink and MATLAB, respectively, to facilitate the attainment of this goal. The formulated strategy has been subjected to empirical testing across a multitude of conditions and scenarios, which are thoroughly elucidated in the ensuing subsections.*

### *4.1. Scenario 1 (Performance of the TRPO-Based NBC Under Normal Configuration and with CTD)*

In order to elucidate the operational properties of the system within a conventional framework frequently represented in the prevailing scholarly literature, the experimental system and methodology are established with the default parameters delineated in Table 1. In recognition of the pivotal role played by the wireless connection in the closed-loop system, a communication time delay (CTD) is integrated to replicate the wireless connectivity. This delay is implemented and calibrated employing both fixed values, [0.005], and stochastic values, [0.006 0.05]. The system's performance is illustrated under standard conditions (controller without CTD approach), a fixed CTD value, and a stochastic CTD value in Figure 6. From outcomes of Figure 6, one can confirm superior level of performance can be achieved by suggested schemes than other designed controllers.

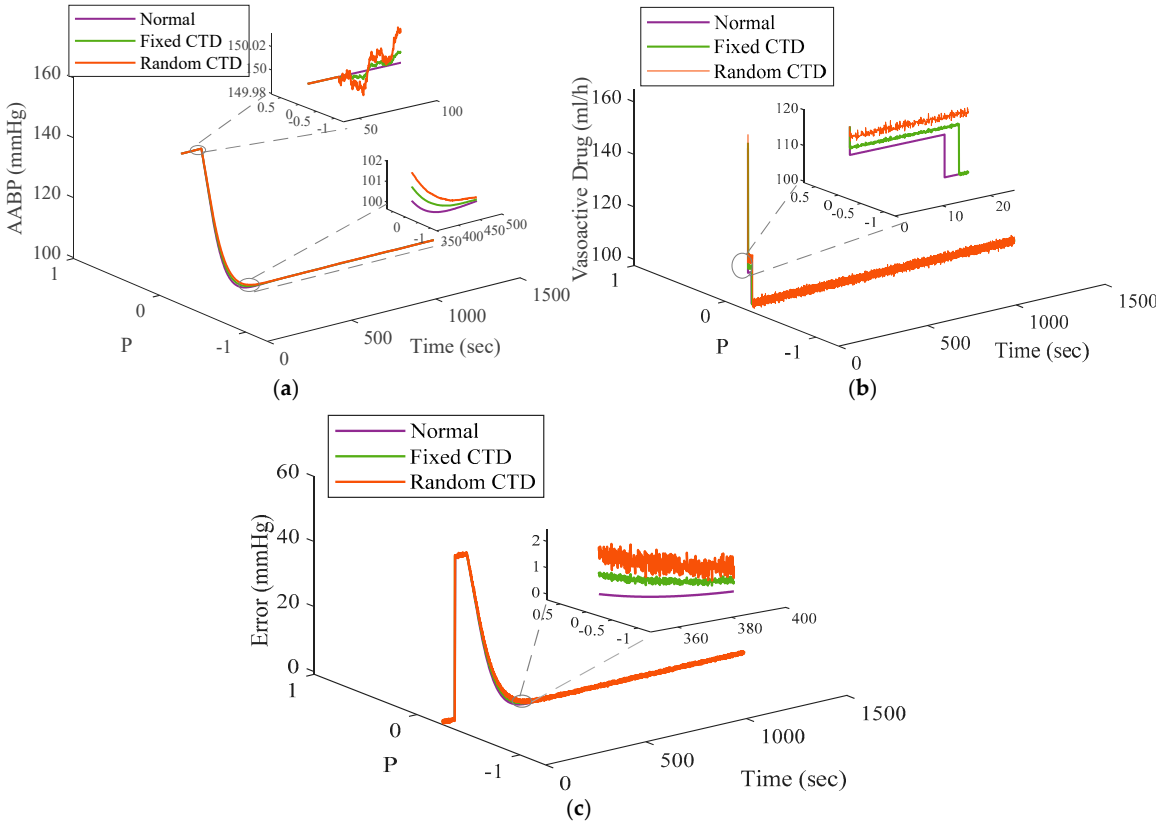



**Figure 6.** The functionality of the system: (**a**) Normal, (**b**) fixed CTD, (**c**) random CTD.

*4.2. Scenario 2 (Performance of the TRPO-Based NBC When Facing Parameters Changing (Different Patients) and Hybrid Noise)*

The outcomes of the strategy employed for AABP level, under P variation across various methodologies (normal, fixed CTD, random CTD) are depicted in Figure 7. Also, the outputs of the system for three mentioned agents under drug action ($X_{i1}$) variation are illustrated in Figure 8. As a final adaptive analysis implemented for the system, the functionality of the closed-loop AABP under reference input changing is pictured in Figure 9.

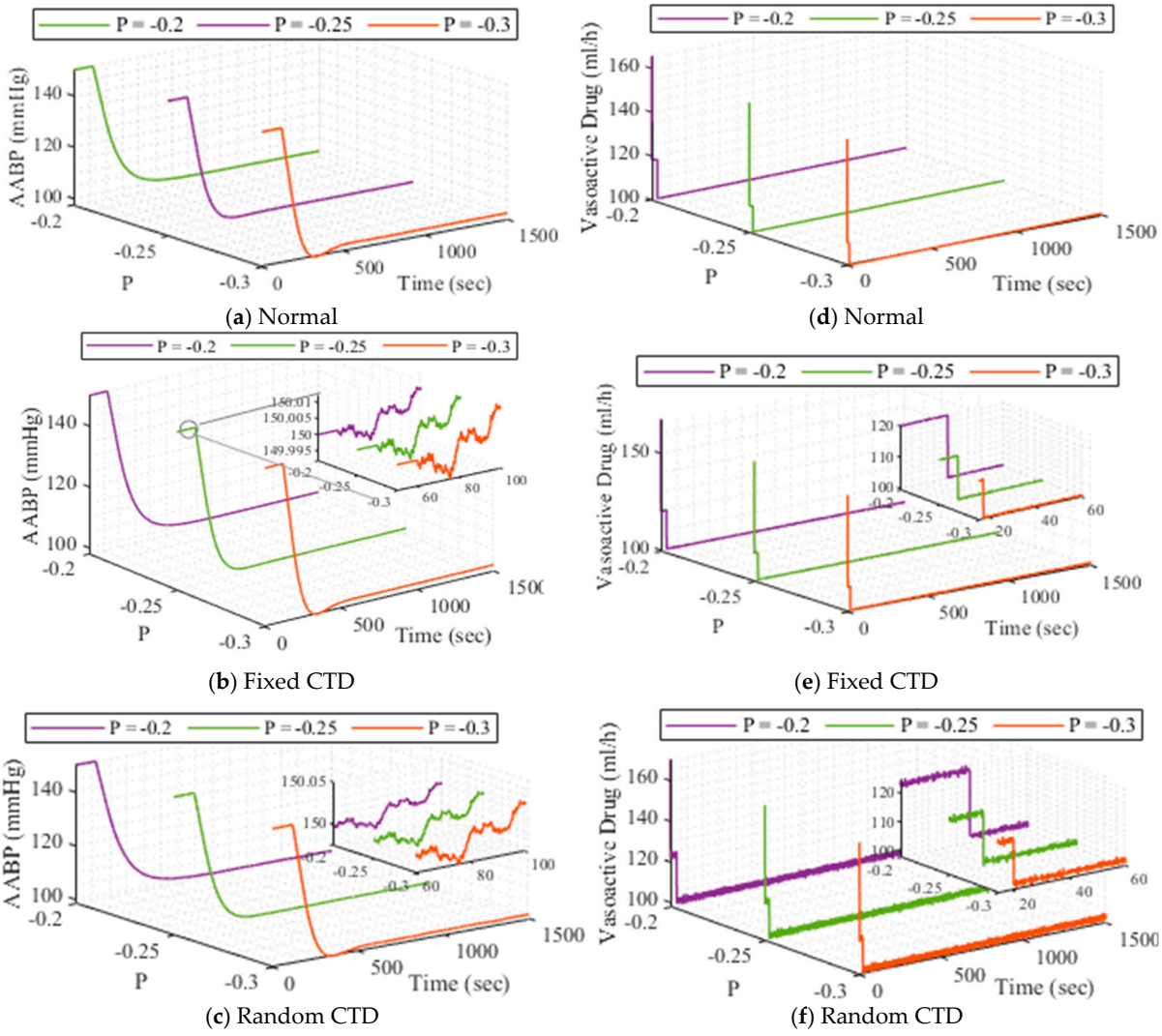



**Figure 7.** The outcomes of the developed strategy when *P* changes, (**a**–**c**) is for AABP, and (**d**–**f**) is for drug dosage.

The simulation results are presented individually for distinct scenarios to illustrate the strategy's adaptability for various patients and the effect of wireless connectivity on system efficiency. The results indicate that the established framework maintains a consistent performance in managing and mitigating the AABP rate through the modulation of SNP. Notwithstanding the influence of the CTD technique on system operations, the system demonstrates stable performance and effectively diminishes and keeps AABP within an acceptable threshold.

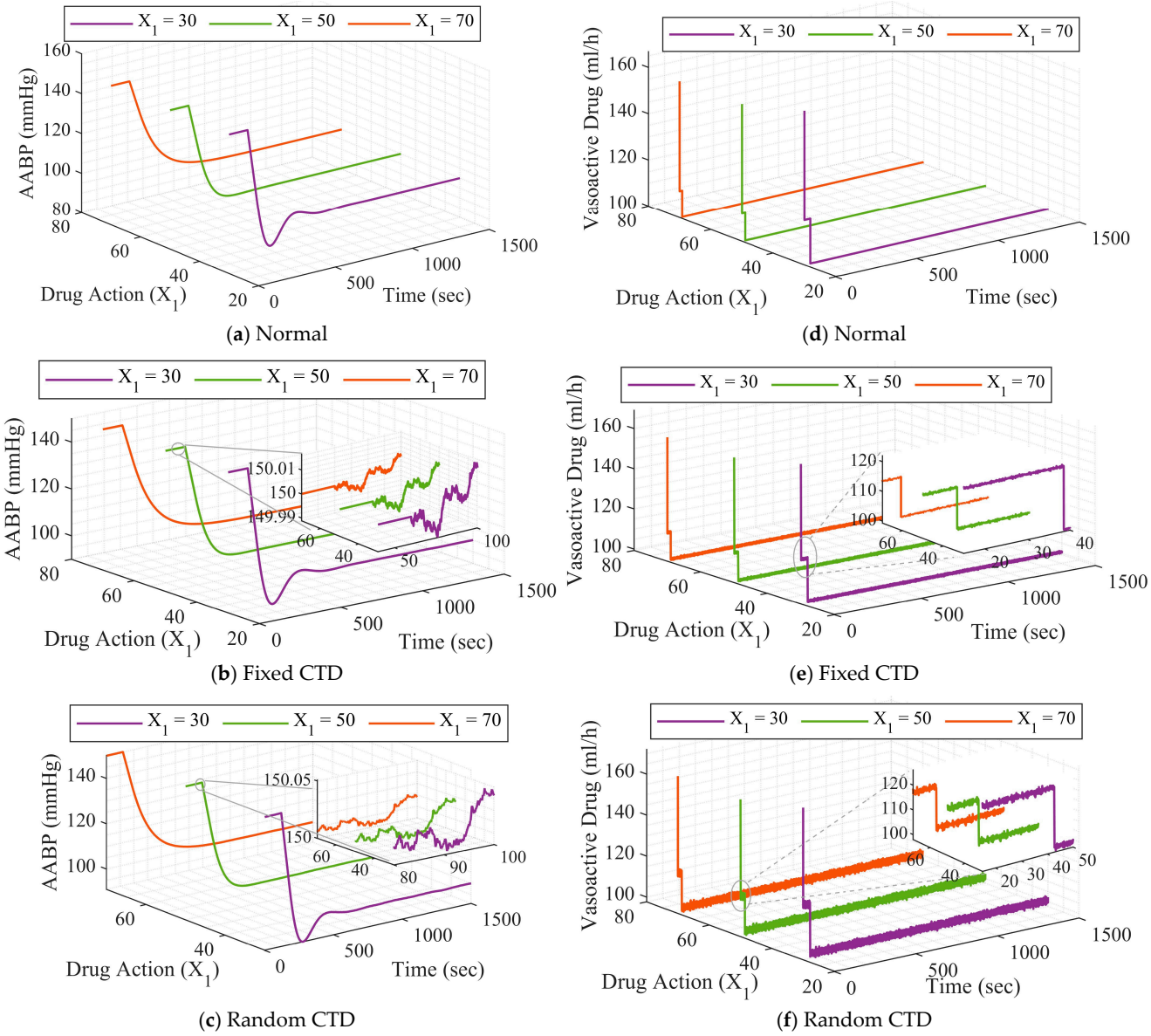



**Figure 8.** The outcomes of the developed strategy when $X_{i1}$ changes, (**a**–**c**) is for AABP, and (**d**–**f**) is for drug dosage.

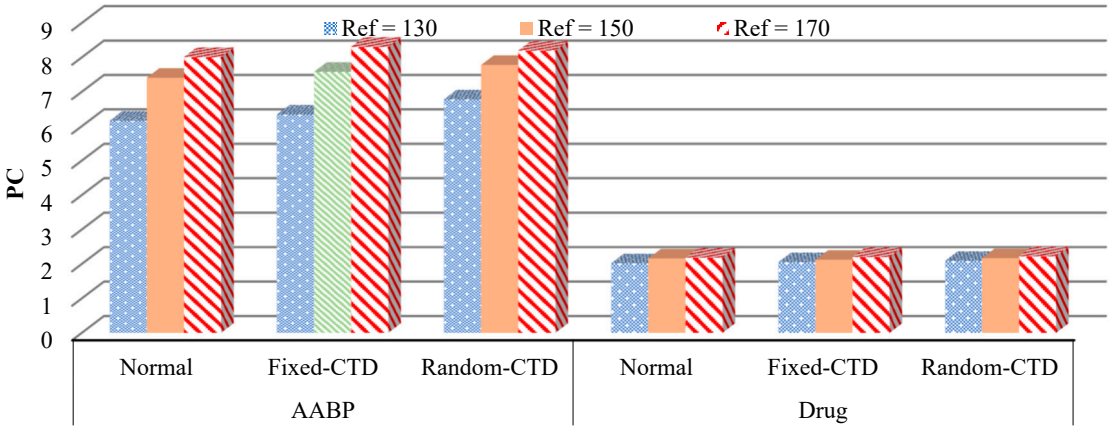



**Figure 9.** The results of the developed strategy when reference changes.

*4.3. Scenario 3 (Quantitative Analysis and Comparative Assessment)*

Following the demonstration of the superior efficacy of the devised strategy through visual representations, a comprehensive statistical analysis for AABP, drug infusion rate, and error was performed by incorporating various metrics for diverse scenarios and conducting comparative studies in two categories: (i) against state-of-the-art methodologies, (ii) in relation to prior research.

To achieve this objective, two specific metrics were used, namely performance criteria (PC) and root mean square (RMS), which were derived using the following expressions.

$$PC = \int_0^{\infty} t.\, agents^2.\, dt \tag{36}$$

$$RMS = \sqrt{mean\ (Agents)} \tag{37}$$

In the interest of elucidation and to mitigate any potential ambiguity, the alternative models that were devised for the purpose of comparative analysis are delineated as follows: **Controller 1:** TRPO-based NBC, **Controller 2:** NBC, **Controller 3:** Conventional Controller. The PC outcomes for the various controllers under P and $X_{i1}$ variation are illustrated by the bar chart in Figures 10 and 11, respectively. In the subsequent phase, the PC value for the various stages of reference changing was computed and compared across different categories, and is documented in Table 2. Subsequently, an evaluation of the efficacy of the newly devised approach under the influence of hybrid noise characterized by a variance of 0.001 was conducted. The RMS results obtained were juxtaposed with alternative control mechanisms and are presented in Figure 12. Also, a comparison with previous studies was conducted and is presented in Table 3 as phase 2 of the comparative study. Based on the findings of the statistical and comparative analysis, it is evident that the devised approach demonstrated superior and consistent performance across diverse scenarios. Moreover, a comparison with alternative methodologies in the various aspects revealed the enhanced functionality of the proposed strategy.

As evidenced in the AABP response plot Figure 10a, both the Conventional Controller and the standalone NBC demonstrate a significant deterioration in their capacity to manage the AABP when subjected to variations in sensitivity, particularly in the context of fixed or random CTD scenarios. Their AABP responses are characterized by extended settling times and substantial deviations from the target reference value. Conversely, the TRPO-based NBC successfully preserves its operational efficiency and stability in regulating the AABP, even amidst the challenges posed by communication delays. The AABP response associated with the TRPO-based NBC consistently aligns closely with the desired reference, exhibiting minimal disruption from the fixed or random CTD conditions.

This assertion is further substantiated by the error plot Figure 10b, which reveals that the TRPO-based NBC consistently exhibits the lowest error metrics across all evaluated scenarios, including those involving CTD. In contrast, the alternative controllers encounter a significant escalation in their error measurements when exposed to communication delays.

The drug (SNP) dosage profile plot Figure 10c additionally underscores the TRPO-based NBC's proficiency in sustaining stable and uniform drug administration, even in the context of CTD. The dosage modifications are maintained within stringent control parameters, thereby ensuring a safer and more comfortable treatment experience for the patient. In sharp contrast, the Conventional Controller and standalone NBC present more erratic and less stable drug dosage profiles under CTD conditions, which may result in suboptimal AABP regulation and raise potential concerns regarding patient safety.

The same behavior is observed from Figure 11, where the functionality of the systems is illustrated under drug action changing.

Moreover, as previously indicated, Figure 12 illustrates the operational characteristics of the controllers in the presence of external uncertainties. It is imperative to note that the system parameters are established as standard, $P = -0.25$ and $X_{i1} = 50$s. It is evident that the influence of external noise adversely affects the efficacy of the controllers; however, the proposed controller adeptly addresses these uncertainties, thereby maintaining superior functionality across all agents in comparison to alternative controllers. The RMS results, Figure 12, prove this superiority.

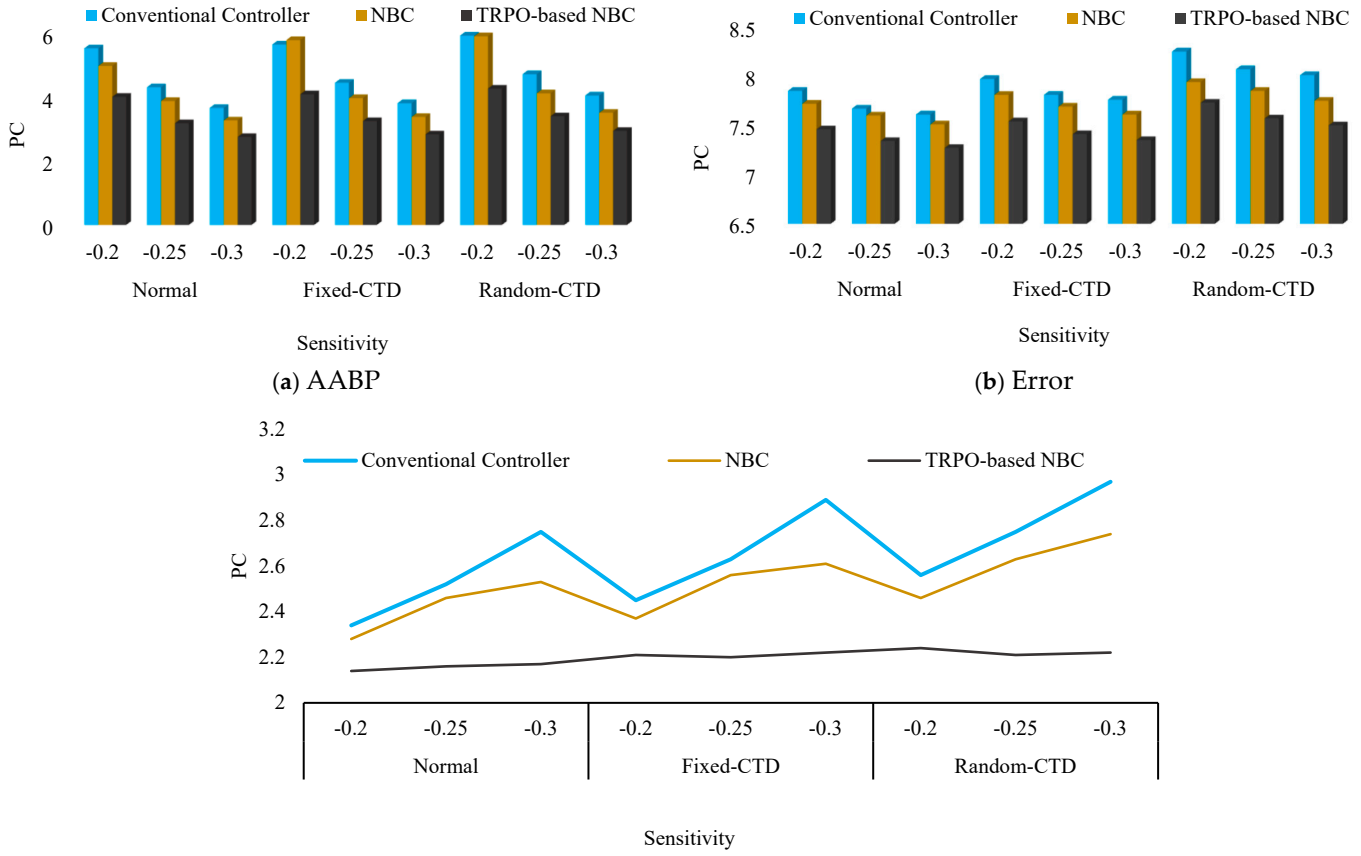



**Figure 10.** The outputs of the various controllers under $P$ changes.

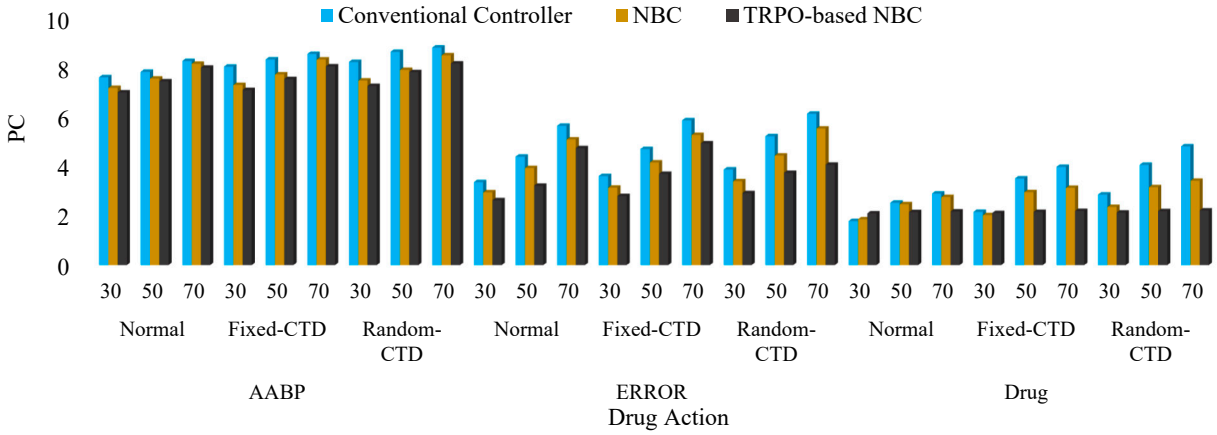



**Figure 11.** The outputs of the various controllers under $X_{i1}$ changes.

**Table 2.** PC results of the controllers under reference input changes.

| Methods | Agents | Normal | | | Fixed CTD | | | Random CTD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **130** | **150** | **170** | **130** | **150** | **170** | **130** | **150** | **170** |
| Controller 1 | AABP | 6.16 | 7.41 | 8.01 | 6.34 | 7.58 | 8.31 | 6.78 | 7.78 | 8.2 |
| | Drug | 2.02 | 2.16 | 2.18 | 2.05 | 2.12 | 2.2 | 2.09 | 2.17 | 2.22 |
| | Error | 2.75 | 3.23 | 4.15 | 3.06 | 3.72 | 4.68 | 3.54 | 3.85 | 4.98 |
| Controller 2 | AABP | 6.48 | 7.56 | 8.38 | 6.75 | 7.76 | 8.72 | 6.97 | 7.93 | 8.59 |
| | Drug | 2.12 | 2.53 | 2.85 | 2.74 | 2.96 | 3.15 | 3.08 | 3.15 | 3.61 |
| | Error | 3.21 | 3.95 | 4.76 | 3.74 | 4.18 | 5.31 | 4.01 | 4.53 | 5.1 |
| Controller 3 | AABP | 6.89 | 7.83 | 8.98 | 7.12 | 8.36 | 9.24 | 7.58 | 8.65 | 8.93 |
| | Drug | 2.28 | 2.56 | 3.02 | 2.96 | 3.52 | 3.95 | 3.54 | 4.05 | 4.58 |
| | Error | 3.72 | 4.41 | 5.19 | 4.23 | 4.73 | 5.85 | 4.65 | 5.27 | 5.81 |

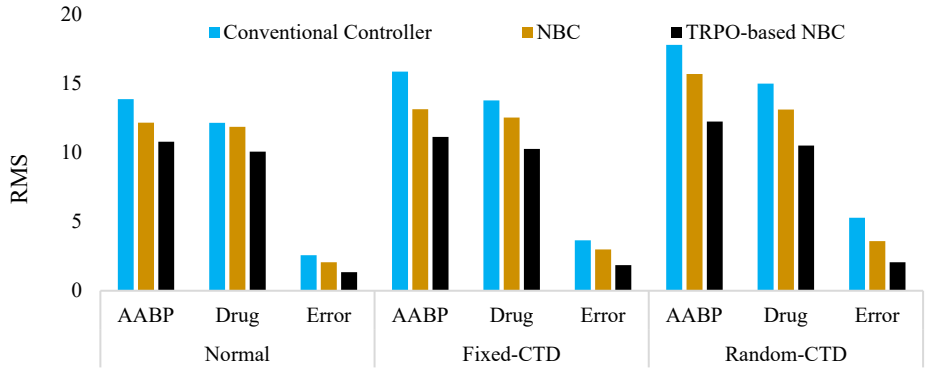



**Figure 12.** RMS results of the created strategy when there is a hybrid noise.

**Table 3.** Comparison of the outcomes of the current research with prior studies.

| Control Strategy | | Setling Time | Over Shoot | IAE |
|---|---|---|---|---|
| IT2-FLC-PID-based Cuckoo Search algorithm [8] | | 350 | 0 | 10,450 |
| Interval Type-2 Fuzzy Logic-based Grey Wolf Optimization (GWO) [9] | | 277 | 0 | 8327 |
| Fractional Order two-layer Fuzzy Logic-based GWO [10] | | 274 | 0 | 8331 |
| ADRC-based CPAG [13] | | 268 | 0 | 8307 |
| 2 degrees of freedom (DOF)-PI-based dragonfly optimization (DA) [45] | | 495 | 0.0951 | 13,124 |
| Type-1 Fuzzy (IT2F) Logic and teaching–learning-based optimization (TLBO) algorithm [46] | | 319 | 0.63 | 9560 |
| Current strategy | Normal | 262 | 0 | 8135 |
| | Fixed CTD | 271 | 0 | 8319 |
| | Random CTD | 282 | 0 | 8346 |

## 5. Conclusions and Forward-Looking Perspectives

The present research initiative has achieved notable advancements in tackling the critical issue of regulating AABP via the modulation of SNP for individuals suffering from BP-related complications, alongside a novel understanding of wireless connectivity in a closed-loop configuration. By employing a closed-loop platform that amalgamates

nonlinear backstepping control (NBC) with Trust Region Policy Optimization (TRPO), the proposed framework exhibits a capacity for real-time dynamic updates of the control system, thereby effectively addressing both AABP and SNP dosage simultaneously. The analysis accentuates the critical role of wireless connectivity in augmenting the operational efficiency of the closed-loop system. The integration of a communication delay network facilitated a comprehensive assessment of the ramifications of wireless connectivity on the system's performance, emphasizing its essential function in ensuring the efficacy of the proposed therapeutic methodology.

The comprehensive experimentation carried out in this investigation, which includes evaluations across a spectrum of patient demographics (varying sensitivity stages, reference inputs, and drug action) and considers hybrid noise and uncertainties with diverse strategies such as normal (without CTD approach), fixed CTD, and random CTD, further highlights the adaptability and dependability of the closed-loop platform. Moreover, the calculation of various metrics such as PC and RMS for the assorted strategies, in conjunction with a comparative analysis with contemporary controllers and prior studies, serves to exemplify the capabilities and superiority of the developed framework through statistical methodologies. These findings pave the way for more personalized and efficient therapeutic interventions for individuals who must manage BP, while concurrently enhancing the closed-loop AABP system for remote patient control. This study endeavors to bridge a knowledge gap and contribute to the scholarly discourse by offering insights that could fortify complex interdisciplinary frameworks for managing the diverse array of BP-related problems. The outcomes of this research could substantially enhance the quality of life for individuals with BP and may also have broader implications for efforts aimed at improving tailored healthcare interventions. Looking forward, the implemented system presents potential applicability to real-world data and clinical analyses, thereby allowing for an exploration of its functionality under practical circumstances.

**Author Contributions:** Conceptualization, M.T.V. and T.H.N.; methodology, M.T.V. and H.L.N.N.T.; software, M.T.V. and S.H.K.; validation, M.T.V., M.R., S.H.K. and T.H.N.; formal analysis, H.L.N.N.T.; investigation, T.H.N. and M.T.V.; resources, M.T.V., S.H.K. and M.R.; data curation, M.T.V. and M.R.; writing—original draft preparation, T.H.N., S.H.K. and M.T.V.; writing—review and editing, M.T.V. and H.L.N.N.T.; supervision, M.T.V.; project administration, T.H.N.; funding acquisition, T.H.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Varon, J.; Marik, P.E. Perioperative hypertension management. *Vasc. Health Risk Manag.* **2008**, *4*, 615–627. [CrossRef] [PubMed]
2. Haas, C.E.; LeBlanc, J.M. Acute postoperative hypertension: A review of therapeutic options. *Am. J. Health-Syst. Pharm.* **2004**, *61*, 1661–1673. [CrossRef] [PubMed]
3. Aronow, W.S. Management of hypertension in patients undergoing surgery. *Ann. Transl. Med.* **2017**, *5*, 227. [CrossRef] [PubMed]
4. Li, X.; Liang, J.; Hu, J.; Ma, L.; Yang, J.; Zhang, A.; On, B.O.T.C. Screening for primary aldosteronism on and off interfering medications. *Endocrine* **2024**, *83*, 178–187. [CrossRef] [PubMed]
5. Bailey, J.M.; Haddad, W.M. Drug dosing control in clinical pharmacology. *IEEE Control Syst. Mag.* **2005**, *25*, 35–51.
6. Wang, X.; Chen, X.; Tang, Y.; Wu, J.; Qin, D.; Yu, L.; Wu, A. The Therapeutic Potential of Plant Polysaccharides in Metabolic Diseases. *Pharmaceuticals* **2022**, *15*, 1329. [CrossRef] [PubMed]

7. Miller, C.P.; Cook, A.M.; Pharm, D.; Case, C.D.; Bernard, A.C. As-needed antihypertensive therapy in surgical patients, why and how: Challenging a paradigm. *Am. Surg.* **2012**, *78*, 250–253. [CrossRef] [PubMed]
8. Dodson, G.M.; Bentley, W.E., IV; Awad, A.; Muntazar, M.; Goldberg, M.E. Isolated perioperative hypertension: Clinical implications & contemporary treatment strategies. *Curr. Hypertens. Rev.* **2014**, *10*, 31–36.
9. Li, H.; Wang, Y.; Fan, R.; Lv, H.; Sun, H.; Xie, H.; Xia, Z. The effects of ferulic acid on the pharmacokinetics of warfarin in rats after biliary drainage. *Drug Des. Dev. Ther.* **2016**, *10*, 2173–2180. [CrossRef]
10. Slate, J.; Sheppard, L.; Rideout, V.; Blackstone, E. A model for design of a blood pressure controller for hypertensive patients. *IFAC Proc. Vol.* **1979**, *12*, 867–874. [CrossRef]
11. Sharma, R.; Deepak, K.; Gaur, P.; Joshi, D. An optimal interval type-2 fuzzy logic control based closed-loop drug administration to regulate the mean arterial blood pressure. *Comput. Methods Programs Biomed.* **2020**, *185*, 105167. [CrossRef] [PubMed]
12. Sharma, R.; Kumar, A. Optimal Interval type-2 fuzzy logic control based closed-loop regulation of mean arterial blood pressure using the controlled drug administration. *IEEE Sens. J.* **2022**, *22*, 7195–7207. [CrossRef]
13. Kumar, A.; Raj, R. Design of a fractional order two layer fuzzy logic controller for drug delivery to regulate blood pressure. *Biomed. Signal Process. Control* **2022**, *78*, 104024. [CrossRef]
14. Ahmadpour, M.R.; Ghadiri, H.; Hajian, S.R. Model predictive control optimisation using the metaheuristic optimisation for blood pressure control. *IET Syst. Biol.* **2021**, *15*, 41–52. [CrossRef] [PubMed]
15. Padmanabhan, R.; Meskin, N.; Haddad, W.M. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed. Signal Process. Control* **2015**, *22*, 54–64. [CrossRef]
16. Mai, V.; Alattas, K.A.; Bouteraa, Y.; Ghaderpour, E.; Mohammadzadeh, A. Personalized Blood Pressure Control by Machine Learning for Remote Patient Monitoring. *IEEE Access* **2024**, *12*, 83994–84004. [CrossRef]
17. da Silva, S.J.; Scardovelli, T.A.; da Silva Boschi, S.R.M.; Rodrigues, S.C.M.; da Silva, A.P. Simple adaptive PI controller development and evaluation for mean arterial pressure regulation. *Res. Biomed. Eng.* **2019**, *35*, 157–165. [CrossRef]
18. Malagutti, N.; Dehghani, A.; Kennedy, R.A. Robust control design for automatic regulation of blood pressure. *IET Control Theory Appl.* **2013**, *7*, 387–396. [CrossRef]
19. Faraji, B.; Gheisarnejad, M.; Esfahani, Z.; Khooban, M.-H. Smart sensor control for rehabilitation in Parkinson's patients. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 267–275. [CrossRef]
20. Berardehi, Z.R.; Zhang, C.; Taheri, M.; Roohi, M.; Khooban, M.H. A Fuzzy Control Strategy to Synchronize Fractional-Order Nonlinear Systems Including Input Saturation. *Int. J. Intell. Syst.* **2023**, *2023*, 1550256. [CrossRef]
21. Faraji, B.; Gheisarnejad, M.; Rouhollahi, K.; Esfahani, Z.; Khooban, M.H. Machine learning approach based on ultra-local model control for treating cancer pain. *IEEE Sens. J.* **2020**, *21*, 8245–8252. [CrossRef]
22. Malik, K.; Tayal, A. Comparison of nature inspired metaheuristic algorithms. *Int. J. Electron. Electr. Eng.* **2014**, *7*, 799–802.
23. Vaidyanathan, S.; Azar, A.T. An introduction to backstepping control. In *Backstepping Control of Nonlinear Dynamical Systems*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 1–32.
24. Bing, P.; Liu, W.; Zhai, Z.; Li, J.; Guo, Z.; Xiang, Y.; Zhu, L. A novel approach for denoising electrocardiogram signals to detect cardiovascular diseases using an efficient hybrid scheme. *Front. Cardiovasc. Med.* **2024**, *11*, 1277123. [CrossRef]
25. Peng, C.C.; Li, Y.; Chen, C.L. A robust integral type backstepping controller design for control of uncertain nonlinear systems subject to disturbance. *Int. J. Innov. Comput. Inf. Control* **2011**, *7*, 2543–2560.
26. Bu, X.; Wang, Q.; Hou, Z.; Qian, W. Data driven control for a class of nonlinear systems with output saturation. *ISA Trans.* **2018**, *81*, 1–7. [CrossRef] [PubMed]
27. Shani, L.; Efroni, Y.; Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5668–5675.
28. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
29. Touati, A.; Zhang, A.; Pineau, J.; Vincent, P. Stable policy optimization via off-policy divergence regulari-zation. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), Virtual, 3–6 August 2020; pp. 1328–1337.
30. Faraji, B.; Rouhollahi, K.; Paghaleh, S.M.; Gheisarnejad, M.; Khooban, M.-H. Adaptive multi symptoms control of Parkinson's disease by deep reinforcement learning. *Biomed. Signal Process. Control* **2023**, *80*, 104410. [CrossRef]
31. Su, Y.; Tian, X.; Gao, R.; Guo, W.; Chen, C.; Chen, C.; Lv, X. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Comput. Biol. Med.* **2022**, *145*, 105409. [CrossRef] [PubMed]
32. Zhang, R.; Li, Z.; Pan, X.; Ma, Z.; Dai, Y.; Mohammadzadeh, A.; Zhang, C. Adaptive Drug Delivery to Control Mean Arterial Blood Pressure by Reinforcement Fuzzy Q-learning. *IEEE Sens. J.* **2024**, *24*, 30968–30977. [CrossRef]
33. Ahmed, S.; Özbay, H. Design of a switched robust control scheme for drug delivery in blood pressure regulation. *IFAC-Pap.* **2016**, *49*, 252–257. [CrossRef]

34. Parvaresh, A.; Abrazeh, S.; Mohseni, S.-R.; Zeitouni, M.J.; Gheisarnejad, M.; Khooban, M.-H. A novel deep learning backstepping controller-based digital twins technology for pitch angle control of variable speed wind turbine. *Designs* **2020**, *4*, 15. [CrossRef]
35. Berardehi, Z.R.; Zhang, C.; Taheri, M.; Roohi, M.; Khooban, M.H. Implementation of TS fuzzy approach for the synchronization and stabilization of non-integer-order complex systems with input saturation at a guaranteed cost. *Trans. Inst. Meas. Control* **2023**, *45*, 2536–2553. [CrossRef]
36. Al Younes, Y.; Drak, A.; Noura, H.; Rabhi, A.; El Hajjaji, A. Nonlinear integral backstepping—Model-free control applied to a quadrotor system. *Proc. Int. Conf. Intell. Unmanned Syst* **2014**, *10*, 1–6.
37. Lele, S.; Gangar, K.; Daftary, H.; Dharkar, D. Stock Market Trading Agent Using On-Policy Reinforcement Learning Algorithms. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3582014 (accessed on 16 December 2024).
38. Zhang, Y.; Ross, K.W. On-policy deep reinforcement learning for the average-reward criterion. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12535–12545.
39. Li, S.E. Deep reinforcement learning. In *Reinforcement Learning for Sequential Decision and Optimal Control*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 365–402.
40. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust region policy optimization. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1889–1897.
41. Roijers, D.M.; Steckelmacher, D.; Nowé, A. Multi-objective reinforcement learning for the expected utility of the return. In Proceedings of the Adaptive and Learning Agents workshop at FAIM, Stockholm, Sweden, 14–15 July 2018; Volume 2018.
42. Sun, M.; Ellis, B.; Mahajan, A.; Devlin, S.; Hofmann, K.; Whiteson, S. Trust-region-free policy optimization for stochastic policies. *arXiv* **2023**, arXiv:2302.07985.
43. Meng, W.; Zheng, Q.; Shi, Y.; Pan, G. An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 2223–2235. [CrossRef] [PubMed]
44. Frei, C.W.; Derighetti, M.; Morari, M.; Glattfelder, A.H.; Zbinden, A.M. Improving regulation of mean arterial blood pressure during anesthesia through estimates of surgery effects. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 1456–1464. [CrossRef]
45. Pachauri, N.; Ghousiya Begum, K. Automatic drug infusion control based on metaheuristic H2 optimal theory for regulating the mean arterial blood pressure. *Asia-Pac. J. Chem. Eng.* **2021**, *16*, e2654. [CrossRef]
46. Karar, M.E.; Mahmoud, A.S.A. Intelligent Networked Control of Vasoactive Drug Infusion for Patients with Uncertain Sensitivity. *Comput. Syst. Sci. Eng.* **2023**, *47*, 721–739. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.