Closed-Loop Deep Brain Stimulation With Reinforcement Learning and Neural Simulation

Chia-Hung Cho, Pin-Jui Huang, Meng-Chao Chen, and Chii-Wann Lin

***Abstract*— Deep Brain Stimulation (DBS) is effective for movement disorders, particularly Parkinson's disease (PD). However, a closed-loop DBS system using reinforcement learning (RL) for automatic parameter tuning, offering enhanced energy efficiency and the effect of thalamus restoration, is yet to be developed for clinical and commercial applications. In this research, we instantiate a basal ganglia-thalamic (BGT) model and design it as an interactive environment suitable for RL models. Four finely tuned RL agents based on different frameworks, namely Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (TD3), Proximal Policy Optimization (PPO), and Advantage Actor-Critic (A2C), are established for further comparison. Within the implemented RL architectures, the optimized TD3 demonstrates a significant 67% reduction in average power dissipation when compared to the open-loop system while preserving the normal response of the simulated BGT circuitry. As a result, our method mitigates thalamic error responses under pathological conditions and prevents overstimulation. In summary, this study introduces a novel approach to implementing an adaptive parameter-tuning closed-loop DBS system. Leveraging the advantages of TD3, our proposed approach holds significant promise for advancing the integration of RL applications into DBS systems, ultimately optimizing therapeutic effects in future clinical trials.**

***Index Terms*— Basal ganglia-thalamic (BGT) network, closed-loop deep brain stimulation (cl-DBS), Parkinson's disease (PD), reinforcement learning (RL).**

Received 16 March 2024; revised 18 August 2024; accepted 16 September 2024. Date of publication 20 September 2024; date of current version 27 September 2024. This work was supported by the Minister of Science and Technology Council, Taiwan, under Grant MOST 111-2221-E-002-079-MY3. *(Corresponding author: Chii-Wann Lin.)*

Chia-Hung Cho is with the Department of Biomedical Engineering, National Taiwan University, Taipei 100, Taiwan (e-mail: r09528018@g.ntu.edu.tw).

Pin-Jui Huang is with the Graduate Degree Program of Artificial Intelligence, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (e-mail: i309505013.eic09g@nctu.edu.tw).

Meng-Chao Chen is with the Department of Biomedical Engineering, National Taiwan University, Taipei 100, Taiwan, and also with the Department of Neurosurgery, China Medical University Hospital, Taipei Branch, Taipei 100, Taiwan (e-mail: neuronxx@gmail.com).

Chii-Wann Lin is with the Department of Biomedical Engineering, National Taiwan University, Taipei 100, Taiwan, also with the Biomedical Technology and Device Laboratories, Industrial Technology Research Institute, Hsinchu 310, Taiwan, and also with the Center for Artificial Intelligence Research, University of Tsukuba, Tsukuba 305-8577, Japan (e-mail: cwlinx@ntu.edu.tw).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2024.3465243, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2024.3465243

I. Introduction

PARKINSON'S disease (PD) is a chronic neurodegenerative disorder affecting the central nervous system. It is cited as the second most prevalent neurodegenerative disease after Alzheimer's [1], affected over 10 million people globally [2]. The degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNc) [3] leads to motor symptoms such as tremors, rigidity, bradykinesia, and postural instability [4], as well as non-motor symptoms including mood changes and swallowing difficulties. While Levodopa/L-dopa is effective in the early stages of PD, its benefits diminish over time, leading to motor complications. High-frequency deep brain stimulation (DBS) ($\geq$100 Hz) offers a promising advanced treatment by regulating activity in targeted brain regions [5]. However, current clinical DBS systems operate in an open-loop regime, which results in higher power consumption, subject-dependent [6], and adverse effects due to overstimulation [7].

Closed-loop deep brain stimulation (cl-DBS) systems are capable of regulating stimulation parameters based on feedback signals and control strateies. Optimizing the cl-DBS algorithm remains crucial for addressing the post-surgical challenge of DBS device [8]. While machine learning (ML) techniques are extensively used in the analysis and prediction of complex systems, deploying new generations of cl-DBS algorithms in live environments remains challenging due to the difficulty of conducting experimentation. Thus, the cl-DBS technique proposed in this study uses physical neural modeling to mimic the fundamental dynamics of the electrophysiological alterations associated with PD, allowing the algorithm to closely replicate the live brain environment and allowing extensive and harmless testing.

In conjunction with the establishment of this environment, development of robust, real-time adaptive algorithms to enhance patient-specific adaptability and address long-term changes in neurological conditions is essential for advancing DBS therapy. Reinforcement learning (RL) has emerged as a powerful technique that enables agents to perceive and interpret interactive environments, followed by determining actions to achieve the most desirable outcomes by maximizing rewards. In other words, integrating RL techniques can facilitate precise, safe (with constrained action range), and personalized adjustments to stimulation parameters, thereby enhancing the effectiveness and reliability of cl-DBS systems.

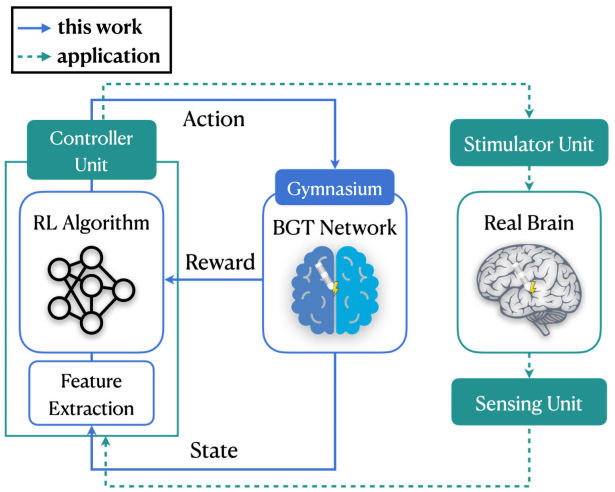

Fig. 1. The overall architecture of this study. The solid blue lines represent what we will implement in this work, whereas the dashed green lines represent a practical direction for the future. Both present the closed-loop characteristics.

Based on a review of current literatures, RL techniques have been increasingly utilized for the treatment of PD via DBS. Lu et al. [9] incorporated a Cerebellar Model Articulation Controller (CMAC) into an actor-critic RL framework, reducing energy consumption by 63.3% compared to open-loop DBS. Krylov et al. [10] used Proximal Policy Optimization (PPO) to train RL agents for suppressing synchronous neuronal activity in models of various oscillations. Gao et al. [11] applied a Markov decision process (MDP) model and convolutional neural networks (CNNs) to alleviate PD symptoms with an average stimulation frequency of 45 Hz. Agarwal et al. [12] used Twin Delayed Deep Deterministic Policy Gradients (TD3) to suppress neuronal synchronization with reduced power consumption, comparing it favorably against other RL algorithms. All these RL models are trained exclusively on pathological (PD state) data, focusing on the alleviation of pathological neuronal activity. However, overlooking the potential coexistence of normal states during training might lead to several issues. Specifically, the model might misinterpret normal data as pathological, resulting in inappropriate stimulation, side effects, suboptimal performance, increased false positives, and potential risks to patient safety. Additionally, within the above articles, feature extraction methods relying on machine learning methods lack explicit guidance on their application to extracellular electrophysiological signals, such as electroencephalograms (EEGs) and local field potentials (LFPs).

In our study, we wrapped the Basal Ganglia-Thalamic (BGT) network that simulate brain dynamics in both normal and pathological states into the Gymnasium [13] environment for developing and comparing RL methods, as depicted in Fig. 1. We prioritize using well-established and validated feature extraction methods for biomarker signals (refer to Section II-B) to ensure their effectiveness in electrophysiological signals during deployment. RL models will explore the best strategy for regulating the frequency and amplitude parameters of DBS. Four mainstream on-policy and off-policy [14] RL frameworks are encompassed in the comparison, namely, Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (TD3), Proximal Policy Optimization (PPO), and Advantage Actor-Critic (A2C). Results demonstrate the effectiveness and superiority of our TD3-based method in terms of power efficiency and mitigation of error response.

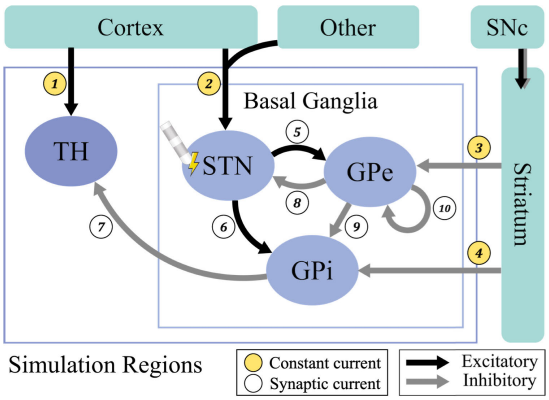

Fig. 2. Illustration of the simulated regions (purple box) and the related currents. Purple ovals are the four neuron types in the basal ganglia-thalamus (BGT) network, containing 10 neurons in each nucleus. Excitatory inputs are represented by black arrows, including ① input from the sensorimotor cortex (I_{SM}), ② constant bias current, $I_{app(STN)}$, to STN, ③ constant bias current, $I_{app(GPe)}$, from Striatum to GPe, ④ constant bias current, $I_{app(GPi)}$, from Striatum to GPi, ⑤ synaptic current from STN to GPe ($I_{STN\rightarrow GPe}$), and ⑥ synaptic current from STN to GPi ($I_{STN\rightarrow GPi}$). Inhibitory inputs are indicated by gray arrows, namely ⑦ synaptic current from GPi to TH ($I_{GPi\rightarrow TH}$), ⑧ synaptic current from GPe to STN (I_{STN}), ⑨ synaptic current from GPe to GPi ($I_{GPe\rightarrow GPi}$), and ⑩ synaptic current from GPe to itself ($I_{GPe\rightarrow GPe}$). Refer to Equation (1),(2), (3).

II. Methods

A. BGT Network Model Simulation

We construct the interactive BGT network based on the Rubin-Terman model [15], [16], [17] focusing on key neural nuclei within the basal ganglia (BG). The subthalamic nucleus (STN), external globus pallidus (GPe), internal globus pallidus (GPi), and thalamus (TH) relay neurons are crucial components in our simulation. Employing conductance-based models, we simulate these four nuclei, interconnected through inhibitory and excitatory synapses (refer to Fig. 2.) Each nucleus comprises 10 neurons to balance fidelity and computational efficiency. The parameters and ordinary differential equations (ODE) of this biophysics model are originated from the work by So et al. [17] and are implemented in Python. Consult the supplementary material for comprehensive understanding of equations and parameters. The BGT network simulation encompasses both normal/healthy and PD/pathological conditions for better RL model generalization.

The membrane potential (v_a) of each neuron obeys Kirchhoff's current balance law, where the subscript a denotes the sub-region, and is presented mainly in differential form as follows:

$$C_m \frac{dv_{STN}}{dt} = -I_L - I_{Na} - I_K - I_T - I_{Ca} - I_{AHP} - I_{GPe\rightarrow STN} + I_{app(STN)} + I_{DBS}, \quad (1)$$

$$C_m \frac{dv_{GPe/i}}{dt} = -I_L - I_{Na} - I_K - I_T - I_{Ca} - I_{AHP} - I_{STN\rightarrow GPe/i} - I_{GPe\rightarrow GPe/i} + I_{app(GPe/i)}, \quad (2)$$

$$C_m \frac{dv_{TH}}{dt} = -I_L - I_{Na} - I_K - I_T - I_{GPi\to TH} + I_{SM}. \quad (3)$$

In the neuronal models, the term $C_m dv/dt$ represents the capacitive current responsible for charging the specific membrane capacitor C_m in STN, GPi, GPe, and TH-type neurons. Currents I_L, I_{Na}, I_K, I_T, I_{Ca}, I_{AHP} correspond to leak, sodium, potassium, low-threshold calcium, high-threshold calcium, and voltage-independent "after hypopolarization" potassium intrinsic ion channel currents. These intrinsic currents are characterized by gating variables that dictate the activation/opening and inactivation/blocking of the channels. External currents, including I_{DBS}, I_{SM}, $I_{\alpha\to\beta}$, and I_{app} (refer to Fig. 2), influence the subsequent elements of the model.

The term I_{DBS} in (1) indicates that the stimulation waveform is directly transmitted to the STN region by the DBS stimulator. Due to safety concerns, I_{DBS} is a symmetric, charge-balanced biphasic pulse, where anodic stimulation comes first and follows the cathodic stimulation with no interphase delay (refer to Fig. 3). Maintaining "charge-balanced" helps prevent undesirable faradic reactions at the electrode-tissue interface over time, which can pose a risk to brain tissue. The pulse width is fixed at 60 μs in consideration of the observed phenomenon that the overall therapeutic window decreases with an increase in the pulse width [18]. Furthermore, fixed pulse width helps minimize charge injection and reduce power consumption [19]. The trained RL agent will intervene in the regulation of additional stimulation parameters, such as frequency and amplitude.

TH neurons do not exhibit intrinsic firing properties without sensorimotor input (I_{SM}). I_{SM} is modeled as a series of anodal, monophasic current pulses with an amplitude of $3.5\mu A/cm^2$ and a pulse duration of $5ms$. The instantaneous frequencies of this pulse conform to a gamma distribution with an average rate of 14 Hz and a variation of 0.2 to emulate the irregular nature of incoming signals from the cortex. As a role of a relay station, TH cells must respond faithfully and promptly to periodic input with a single action potential (AP) [17]. Subsequently, this signal will be transmitted to the brainstem and spinal cord to facilitate the execution of motions. Relay error exhibits a high correlation with motor symptoms, as indicated in [20]. It functions as a quantitative metric for assessing the degree of PD pathology in our study. We quantify the degree of response error using the Error Index (EI), which is formalized as:

$$EI = \frac{N_{error}}{N_{SM}}. \quad (4)$$

According to the equation, EI is defined as the number of error transmissions (N_{error}) over the total number of sensorimotor inputs (N_{SM}). It depends upon the average of all (10) TH channels/neurons. Higher EI indicates a greater dominance of PD in the current circuit and lower relay reliability (RR) of TH neurons.

Currents in the form of $I_{\alpha\to\beta}$ stand for synaptic inhibitory or excitatory current from presynaptic nucleus α ($\alpha \in$ {STN, GPe, GPi}) to postsynaptic nucleus β ($\beta \in$ {GPe, GPi, TH}). According to [17], each STN neuron receives inhibitory input from two GPe neurons. Each GPe or GPi neuron receives excitatory input from two STN neurons and inhibitory input from two other GPe neurons. Finally, each GPi cell projects to one TH cell. In other words, the effect of the overall BG network and external DBS is propagated to TH through GPi, i.e., $I_{GPi\to TH}$, allowing us to evaluate the efficacy of stimulation through the quantified EI.

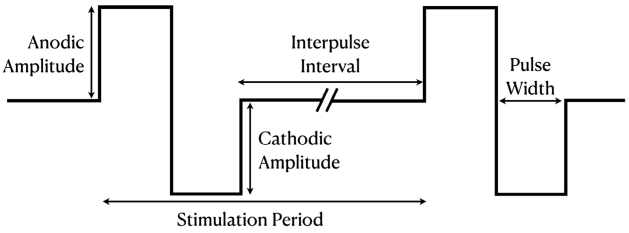

Fig. 3. Illustration of the biphasic, charge-balanced, symmetric DBS pulses we applied throughout our simulation work.

I_{app} denotes the constant external applied/bias currents in STN, GPe, and GPi nuclei, which is the main difference between the healthy and PD states in simulation. Based on the PD etiology, decreased I_{app} level elucidates the effect of insufficient dopamine secretion by SNc since currents from other brain regions or striatum are correspondingly lessened. We apply additional noise (refer to supplementary material) to the I_{app} of GPe neurons to simulate the variability in this variable due to different PD salience in the current circuit.

B. Biomarker Selection

In the BGT network, we call for a discriminative signal as the environmental output. Varied relay properties in the TH neuron are influenced by the $I_{GPi\to TH}$ synaptic current that carries distinct signal representations. $I_{GPi\to TH}$ is comprised of: $I_{GPi\to TH} = g_{GPi\to TH}[v_{TH} - E_{GPi\to TH}]\sum S_{GPi}$, where $g_{GPi\to TH}$ is the maximal synaptic conductance, S_{GPi} denotes the synaptic variable from the presynaptic structure GPi, and $E_{GPi\to TH}$ is the reversal potential across synapses. Among these components, we refer to the synaptic variable-based control strategy proposed by Gorzelic et al. [21], setting S_{GPi} as a biomarker signal. We further examine the correlation between the S_{GPi} signals and TH membrane potentials in three different states in Section III-A.

C. Problem Formulation

We wrapped the BGT network into a customized interactive environment based on Gymnasium ([13], [22]) architecture, devising a tailored interface with appropriate action space, state space, reward function, episode configuration, and step length. As an initial condition, the environment randomly assigns a state from healthy and PD when an episode starts, mimicking the irregular occurrence of PD.

1) Action: Action space comprises the DBS frequency and amplitude value in a total dimension of 2. These values serve as the output of the RL model, while the input to the BGT environment. Studies have evaluated the effects of variation in the DBS parameters and suggested suitable ranges ([18], [19], [23], [24]). Both frequency and amplitude are continuous variables within the range of 100∼185 Hz and 0∼5000 $\mu A/cm^2$, while the pulse width remains fixed at 60 μs. However, these

actions will be set in a normalized range $[-1, 1]$, aligning with the common practice in many RL algorithms that utilize a Gaussian distribution (initially centered at 0 with a standard deviation of 1) for continuous actions. The actual frequency and amplitude value will be denormalized back to the desired range within the BGT environment, by using:

$$\text{actual value} = \frac{(\text{normalized} + 1)}{2} \times (\max - \min) + \min . \quad (5)$$

2) State: The state space comprises the feature extraction value extracted from the biomarker signal S_{GPi} (as detailed in Section II-B) in a total dimension of 6. Contrary to the action space, state values are the input to the RL model, while the output of the BGT environment. Furthermore, all features were normalized by the max–min normalization technique. Extracellular-based feature extraction techniques are as follows:

- Signal standard deviation.
- Hjorth Parameters: Hjorth Parameters, comprising activity, mobility, and complexity, offer a statistical characterization of time-domain signals. Initially developed for EEG analysis due to low computational complexity [25], they have proven effective in enhancing PD diagnosis with an accuracy of up to 89.3% [26]. Calculation defined as:

$$\text{Activity: } A = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2, \quad (6)$$

$$\text{Mobility: } M = \sqrt{\frac{A(\frac{dx_i}{dt})}{A}}, \quad (7)$$

$$\text{Complexity: } C = \frac{M(\frac{dx_i}{dt})}{M}, \quad (8)$$

  where x_i represents the signal values, $\overline{x}$ is the mean of the signal, and N is the number of samples.
- Beta Band Power: Increasing evidence indicates a correlation between beta-frequency band (12)–30 Hz) oscillation powers in the LFPs recorded in the STN of PD patients and motor impairments such as bradykinesia and rigidity [27]. PD patients exhibit elevated beta power spectra in both STN and GPi neurons, which can be suppressed by adequate stimulation amplitude or medication. Calculation defined as:

$$S_x(f) = \frac{1}{T}|X(f)|, \quad (9)$$

$$P_\beta = \int_{f_{low}}^{f_{high}} S_x(f)\, df, \quad (10)$$

  where $X(f)$ is the Fourier transform of the filtered signal x_β, T is the total time duration of the signal, $S_x(f)$ is the power spectral density of x_β, f_{low} and f_{high} are cutoff frequencies for the bandpass filter, which is 12 and 30 in our case, and finally the P_β is the desired beta-band power.
- Sample Entropy (SampEn): SampEn has proven effective in evaluating the complexity of physiological time-series signals and diagnosing disease states [28]. Its advantages over approximate entropy (ApEn), such as data length independence and ease of implementation, make it a preferable choice. Lower sample entropy values indicate a higher degree of self-similarity in the dataset, reflecting lower complexity and irregularity, which is often observed in PD cases. In the context of subthalamic nucleus-local field potential (STN-LFP) signals, neuronal entropy exhibited a progressive increase with the rise of DBS amplitude, coinciding with the suppression of beta-band oscillation—a characteristic that can be interpreted as an inverse indicator [29].
  Formula defined as:

$$SampEn(m, r, N) = -\ln \frac{C(m+1, r)}{C(m, r)}, \quad (11)$$

  where N is the data length, m is the embedding dimension (default = 2), r is the radius of the neighborhood (default = $0.2 \times std(x_i)$), $C(m+1, r)$ is the number of embedded vectors of length $m + 1$ having a Chebyshev distance inferior to r, and $C(m, r)$ is the number of embedded vectors of length m having a Chebyshev distance inferior to r. Sample entropy measures the likelihood that vectors of length m that are close to each other will remain close when their length increases to $m + 1$.

To ensure the applicability of these feature extraction methods in real data, we validated them in the EEG dataset from [30] using channels located above the primary motor cortex (C3, FC3, CP3, C5, FC4, C4, C6, CP4). The diagram is presented in the Results III-B.

3) Step Length: The step length significantly influences the time resolution of the action and information content of the state space, presenting a trade-off. A shorter step length provides higher resolution in the control action space and more dynamic DBS waveforms. However, this comes at the cost of potentially diminishing the meaningfulness of state signals to the RL agent and limiting the observation of long-term features. In our study, we selected a 100-millisecond (ms) step length, guaranteeing the occurrence of at least one ISM input pulse at 14 Hz.

4) Episode Termination Prerequisites: Determining when an episode is done in RL environment depends on the specific context and goals of the task. For a DBS parameter tuning environment, the following criteria should be met:

- EI of the current state is zero (no error response in current state).
- The average EI is below 0.1.
- The average beta band power is suppressed below a threshold value ($\mathcal{T}_\beta$).

Satisfying the above demands will lead to an episode termination, indicating convergence of the episode.

5) Reward: In our design, the reward function combined different aspects into a single, balanced reward function as follows:

$$Reward = R(t) = \alpha \cdot r_1 + \beta \cdot r_2 + \gamma \cdot r_3 + \delta \cdot r_4, \quad (12)$$

where r_1, r_2, r_3, r_4 is respectively identified as "improvement score," "energy consumption," "side effect score," and "compensation score." Positive weighting coefficients (α, δ) imply encouragement, while negatives (β, γ) are for penalty.

Additionally, each component of the reward function is scaled in the range [0, 1] to avoid skewed learning of the RL model.

Crafting the improvement score based on reliability components (EI) can be a beneficial approach, considering that reducing thalamus EI is one of the primary objectives of this task. We define the improvement degree of EI before (EI_{t-1}) and after (EI_t) the action ($I_{DBS}(t)$) as the first reward component with $\alpha = 1.2$:

$$r_1 = EI_{t-1} - EI_t. \tag{13}$$

Next, the energy consumption is calculated using the root mean square of $I_{DBS}(t)$, where the frequency and amplitude components are actions output from the RL model, as:

$$r_2 = I_{RMS}$$
$$I_{RMS} = \sqrt{\frac{1}{T}\int_0^T I_{DBS}^2(t)dt}, \tag{14}$$

where T denotes the duration of the I_{DBS} stimulation on STN neurons, and weighting factor $\beta = -0.8$. r_2 is further scaled by the highest possible value of I_{RMS}, which is based on the upper bound of DBS frequency: 185 Hz and amplitude: 5000 $\mu A/cm^2$.

To expedite the agent to achieve the episode termination goal without intentionally prolonging the episode, we design the "side effect score" as the EI of the current state with $\gamma = -0.5$:

$$r_3 = EI_t \tag{15}$$

Follow with a compensation value for switching off the DBS (zero amplitude) in healthy states for encouragement of energy conservation, with weighting factor $\delta = 0.5$:

$$r_4 = \begin{cases} 1, & \text{if} \quad r_1 \cap r_2 \cap r_3 = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Jointly, the final reward function is normalized as: $R'(t) = R(t)/(\alpha + \beta + \gamma + \delta)$ within $[-1, 1]$ to stabilize training.

D. RL Actor-Critic Frameworks Implementation

In this study, we evaluate the BGT environment using the Soft Actor-Critic (SAC [31]), Twin Delayed Deep Deterministic Policy Gradient (TD3 [32]), Proximal Policy Optimization (PPO [33]), and Advantage Actor-Critic (A2C [34]) frameworks. All models share the same critic and actor architecture, implemented using PyTorch [35].

SAC is an off-policy actor-critic algorithm that incorporates an entropy regularization term for exploration encouragement. Its objective function combines expected return and policy distribution entropy, preventing excessive determinism for improved exploration. The learnable temperature parameter (α), updated through gradient descent, controls entropy regularization strength. Critic and target critic networks guide policy optimization, with soft updates ensuring gradual adaptation. The actor-network employs a Gaussian policy parameterized by the mean and standard deviation for stochasticity.

TD3 addresses issues in deep deterministic policy gradient (DDPG [36]) by reducing the overestimation bias with twin critic networks, delayed updates of the actor, and action noise regularization. It is an off-policy algorithm, similar to SAC, and it leverages the advantages of a replay buffer. This approach enhances data efficiency, diminishes correlations between consecutive samples, facilitates efficient batch learning, and enables the algorithm to revisit and learn from past experiences. The critic networks are updated to minimize the temporal difference (TD) between the predicted Q-values and the target values, in both TD3 and SAC.

PPO is an on-policy algorithm, meaning it learns from the data collected by the current policy. The rollout buffer stores on-policy experiences sampled from the most recent policy to ensure that the learning process remains focused on the current policy. It involves replacing the intricate constrained optimization step in the Trust Region Policy Optimization (TRPO [37]) with a simpler surrogate objective function that incorporates advantage, a clipping mechanism, and the entropy of the policy.

A2C is an on-policy algorithm that integrates policy and value learning, ensuring simplicity and stability in training with synchronous updates. It directly optimizes the policy using the advantage function with the value function baseline, represented as the difference between the estimated value function and the value of the current state. Notably, A2C does not explicitly enforce a trust region constraint, allowing for potentially larger policy updates.

III. Results

A. Environment Simulation Results

Fig. 4 report the voltage traces of TH neuron, the synaptic signal, S_{GPi}, and its scalogram in normal, PD without DBS, and PD with 130 Hz DBS condition. Scalograms are calculated through continuous wavelet transform with the Morse wavelet. During PD state, substantial synchronous in the GPi nuclei is sufficient to affect thalamic activity through large synchronous oscillations/fluctuations in the S_{GPi} signal, resulting in a higher EI compared to other conditions. There is also a substantial difference between the pathological state and the others in the scalogram, presenting the frequency band (10)–20 Hz) of the synchronous neuronal activity. The applied DBS could suppress the oscillating characteristic of S_{GPi}, exhibiting a reduction in the error response in TH neurons to I_{SM} and the band power.

B. Feature Verification

Fig. 5 demonstrates the results of the feature extraction (mentioned in II-C.2) for both synaptic signal (S_{GPi}) from the BGT environment and the EEG dataset from [30] across PD and Healthy Control (HC) participants. In the 64-channel montage of EEG electrodes, the channels most commonly related to PD for further signal analysis are typically those covering the motor cortex and supplementary motor areas. In total, eight channels are selected, including C3, FC3, CP3, C5 within the left central lobe, and C4, FC4, C6, CP4 within the right central lobe. Features are normalized by the max–min normalization technique. The observed consistency in trends between PD and HC states suggests promising potential for their application in

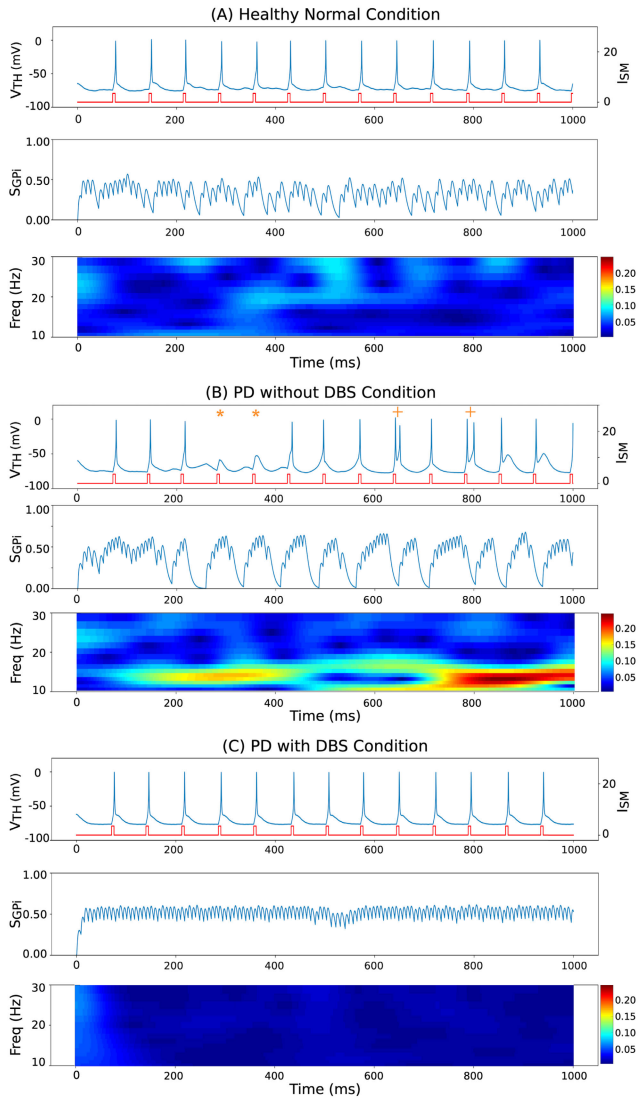

Fig. 4. Thalamus voltage traces, synaptic input signals from GPi to TH (S_{GPi}), and scalogram within the beta band in three conditions: (A) normal/healthy (EI=0.0), (B) PD without DBS (EI=0.5), and (C) PD with DBS conditions(EI=0.0). I_{SM} inputs are highlighted in red pulse. +: represents a "bursting" error response (generating more than one AP); *: represents a "missing" error response (TH neuron signal does not constitute an AP). There is a bright band (high magnitude/power) between 10–20 Hz in (B) PD condition, which is the so-called beta band oscillation. The oscillation is obscure in (A) healthy conditions and is eliminated with biphasic DBS in (C).

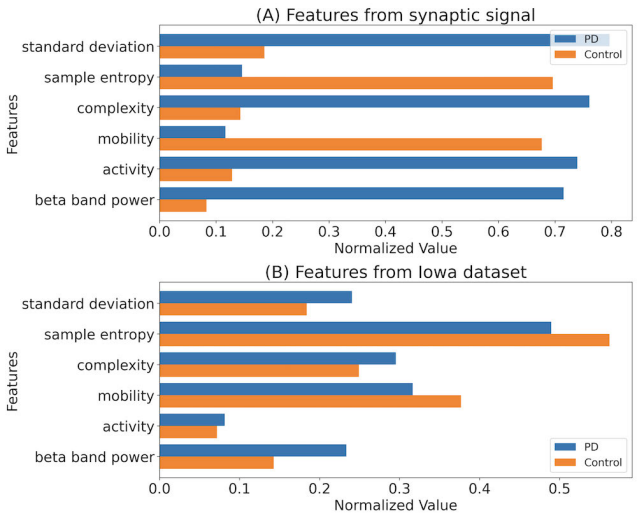

Fig. 5. Observations on the effect of feature extraction in (A) synaptic signals and (B) EEG signals. Eight channels (C3, FC3, CP3, C5, FC4, C4, C6, CP4) are selected from the Iowa dataset in [30] for feature verification.

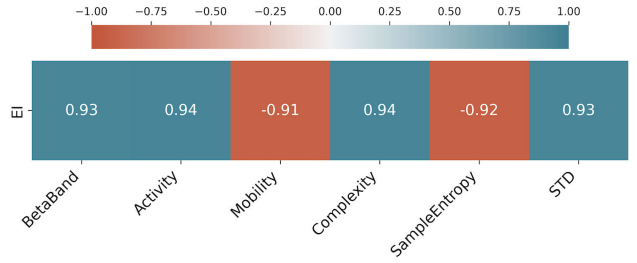

Fig. 6. Pearson correlation coefficients between each feature and the error index (EI) of the TH neurons.

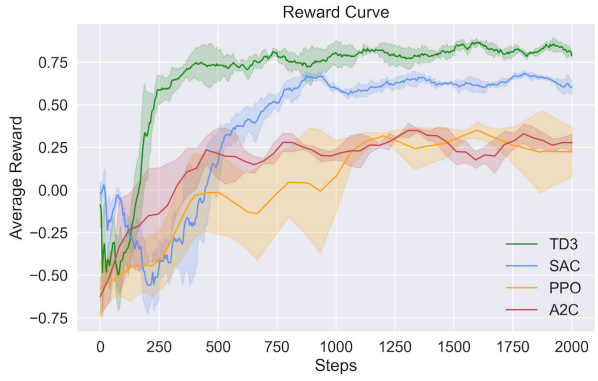

Fig. 7. The reward curve of the RL models across different architectures, including Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (TD3), Proximal Policy Optimization (PPO), and Advantage Actor-Critic (A2C). The x-axis represents the training steps, while the y-axis shows the average rollout reward obtained by the RL models. The shaded regions around each curve represent the ± 1 standard deviation of the rewards, reflecting the variability in the model's performance across training sessions.

subsequent agent deployments. Furthermore, the correlation between each feature and the EI of TH neurons is shown in Fig. 6. All features suggested highly correlation with the EI of TH neurons, allowing the biomarker S_{GPi} to be nicely represented.

C. RL Experimental Results

The reward curves in Fig. 7 for each architecture portray varying levels of performance over each training session. The plot reveals that the TD3 architecture (green line) converges the fastest, stabilizing after just 400 steps, and achieves the highest reward value, demonstrating its efficiency and effectiveness in optimizing stimulation parameters. The SAC model (blue line) converges at a slightly slower rate, attaining the second-highest peak reward. PPO (orange line) and A2C (red line) exhibit nearly identical convergence rewards, with A2C securing a marginally faster convergence time than PPO, indicating a modest performance advantage. The shaded regions surrounding each curve represent the ± 1 standard deviation of the rewards, providing insight into the variability and consistency of each model's performance across episodes. These

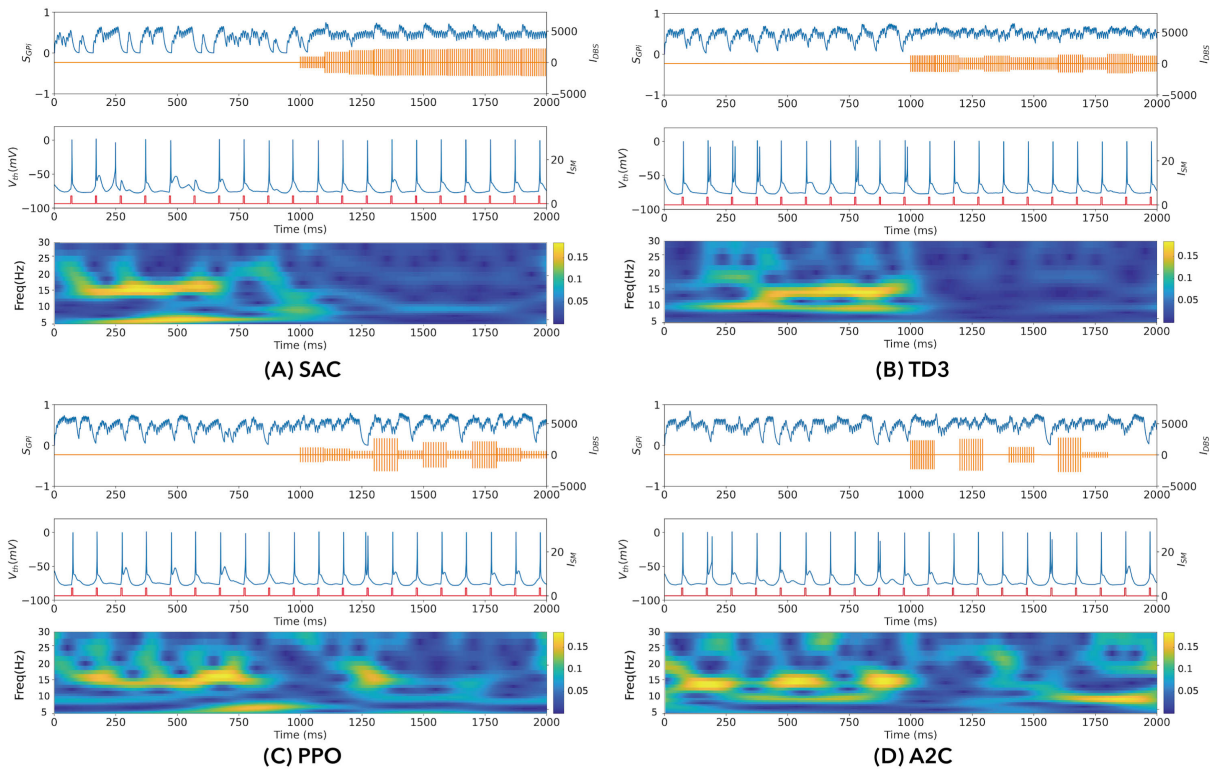

Fig. 8. Control strategy in the **PD** state by (A) SAC, (B) TD3, (C) PPO, and (D) A2C RL agents. Stimulation is activated after 1000 milliseconds. Each subplot includes the biomarker signal (S_{GPi}), action signal (I_{DBS}), thalamus action potentials, sensorimotor input (I_{SM}), and the scalogram of the S_{GPi} signal in the beta frequency band, from top to bottom.

areas emphasize the stability of TD3's superior performance and the slightly greater variability observed in SAC.

Fig. 8 and Fig. 9 illustrate the control strategies performed by agents trained using the SAC, TD3, PPO, and A2C RL frameworks in the PD and healthy state. DBS is activated after 1000 milliseconds (ms). Each subplot includes the biomarker signal (S_{GPi}), action signal (I_{DBS}), thalamus action potentials in response to sensorimotor input (I_{SM}), and the scalogram of the S_{GPi} signal in the beta frequency band. Table I summarizes the quantitative reductions in percentage and average EI for each framework compared to ol-DBS.

In the PD state, the agents are anticipated to administer optimal stimulation based on signal features, effectively mitigating the existing pathology without undue energy expenditure. Fig. 8 reveals that both SAC and TD3 agents manifest actions with low variability, contributing to significant corrections in thalamic relay reliability (both with EI values of 0) and the suppression of oscillations in the beta frequency band. Notably, TD3 exhibits superior energy efficiency compared to SAC. However, under the parameter control of the (on-policy) PPO and A2C agents, the resulting actions show increased variability, and the parameter adjustments lead to less effective suppression in the PD state. Due to the limited suppression effect on the beta band oscillation, a distinct bright band continues to appear in the scalogram after 1000 ms. Quantitatively, the EI values are notably higher, reaching 0.15 and 0.23, respectively, as shown in Table I.

In the healthy control state, guided by the reward design, the agents are expected to minimize or deactivate stimulation to conserve energy without inducing side effects. SAC maintains

TABLE I
EVALUATION METRICS FOR ALL TRAINED RL AGENTS, OPEN-LOOP DBS (OL-DBS), WITHOUT DBS IN BOTH PD AND HEALTHY CONTROL CONDITIONS

		PD state		Healthy Control state	
		Reduced Percentage	Avg. EI	Reduced Percentage	Avg. EI
Off-policy	SAC	58%	0.0	72%	0.0
	TD3	**67%**	**0.0**	**100%**	**0.0**
On-policy	PPO	65%	0.15	78%	0.0
	A2C	62%	0.23	77%	0.01
ol-DBS		-	0.043	-	0.27
Without DBS		-	0.5	-	0.01

a stable output with small amplitude, and the application of stimulation does not result in side effects or an increase in EI. Remarkably, under the TD3 agent's control, it effectively modulates the amplitude to zero, indicating the cessation of stimulation. This control strategy demonstrates significant effectiveness. PPO and A2C strategies typically show higher variability. Although they exhibit stability in mild oscillations in the healthy state, a slightly increased power in the beta frequency band is observed on the scalogram compared to the former two strategies. Their energy efficiency is slightly lower, with values of 78% and 77%, respectively, subsequent to TD3.

Table I shows the quantitative comparison of four trained RL agents, open-loop DBS, and the case without DBS intervention in both PD and healthy control states. In ol-DBS regime, we assume $I_{DBS}(t)$ delivers pulses with frequency

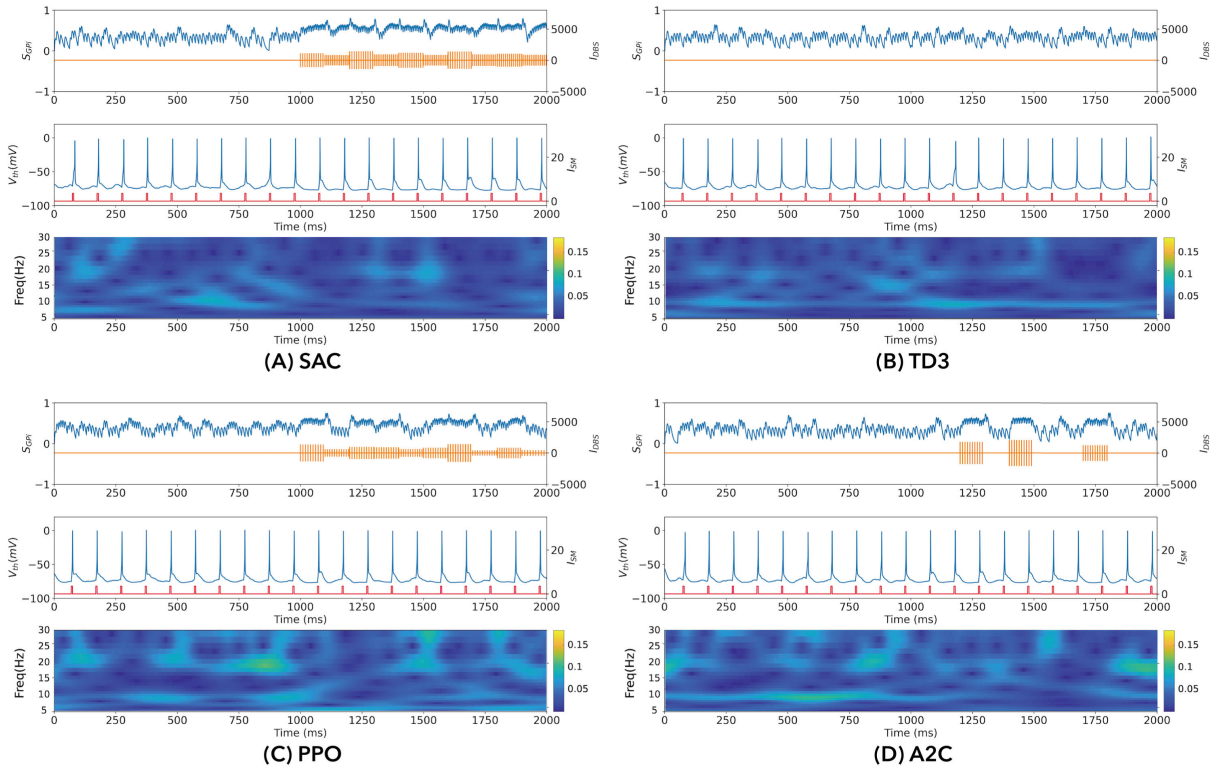

Fig. 9. Control strategy in the **healthy control** state by (A) SAC, (B) TD3, (C) PPO, and (D) A2C RL agents. Stimulation is activated after 1000 milliseconds. Each subplot includes the biomarker signal (S_{GPi}), action signal (I_{DBS}), thalamus action potentials, sensorimotor input (I_{SM}), and the scalogram of the S_{GPi} signal in the beta frequency band, from top to bottom.

of 130 Hz and amplitude of 2500 $\mu A/cm^2$. The reduced percentage is calculated by: $1-(\hat{I_{RMS}}/\acute{I_{RMS}})\times 100\%$, where $\hat{I_{RMS}}$ is for root mean square of target $I_{DBS}(t)$, and $\acute{I_{RMS}}$ is for the corresponding value in ol-DBS. Notably, the elevation of EI in ol-DBS regime under healthy state highlights potential concerns related to overstimulation and its associated side effects, while the restorative effect is constrained in the PD state.

In summary, off-policy approaches exhibit better stability in generating actions for this task and demonstrate superior restoration capability compared to on-policy agents. However, SAC tends to employ a more greedy strategy, resulting in relatively higher energy expenditure. Among the off-policy frameworks, PPO slightly outperforms A2C, with its control strategy resembling SAC in the healthy state. TD3 stands out in both scenarios across all frameworks: in the PD state, it effectively restores thalamic relay reliability, suppresses beta frequency oscillations, and maintains efficient energy usage; in the healthy condition, it conserves energy by deactivating stimulation, preventing side effects.

D. Continuous Episode Evaluation

The outperformed TD3 architecture is further evaluated in continuous episodes to mimic real-world deployment. Conterminous states with RL model intervention and its corresponding beta band scalogram are shown in Fig. 10. We extract the information of the PD occurrence index from the environment to portray the "ground truth" of a healthy control or parkinsonism state for reference. Between 0 and 200 ms, as a healthy state, the model maintains a stimulation amplitude of zero. After 200 ms, with the onset of PD at 200 ms and 400 ms, a prominent bright beta band appears in the S_{GPi} scalogram, accompanied by bursting error in the TH neurons. The model then applies stimulation to effectively suppress beta band power between 300 and 400 ms. The normal state reemerges at 600 ms, leading to a slight application of DBS amplitude. A new bright beta band appears between 800 and 1000 ms, prompting the RL model to take an additional step to better detect the PD occurrence. DBS is then applied again after 1000 ms to suppress this beta band. This showcases the model's capability to adapt and respond to dynamic conditions in a continuous, real-world setting.

IV. Discussion

Our study demonstrated the improvement of cl-DBS systems via RL-based architectures compared to existing systems. This finding underscores the potential of RL to offer more precise and adaptive treatment by automatically adjusting stimulation parameters based on environmental feedback. Within the interactive environment, we focused on the core dynamic changes of action potentials of neuron cells in the BGT network. This simulation allows to capture more precise information between cell synapses. Additionally, unlike earlier studies, the mechanism with intermittent pathological and healthy states aims to enhance the robustness of our RL model across varying neuronal conditions.

We fine-tuned parameters in mainstreamed RL architectures in evaluation for energy efficiency and error correction. Due to shared dynamics between PD and healthy states in our environment, off-policy algorithms efficiently reuse

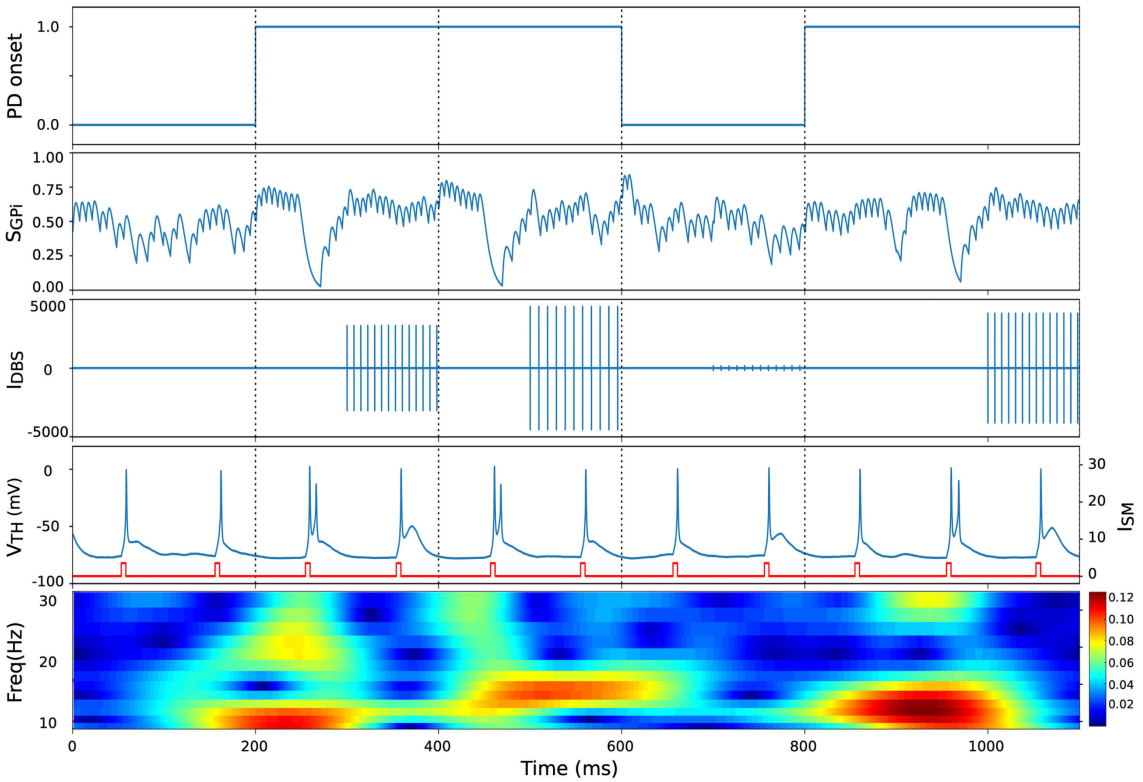

Fig. 10. Continuous episode with RL agent intervention and its corresponding beta band scalogram. The dotted line separates each episode. Frequency of 132 Hz and amplitude of 3087 $\mu A/cm^2$ at 300∼400 ms; 104 Hz and 4473 $\mu A/cm^2$ at 500∼600 ms; 116 Hz and 150 $\mu A/cm^2$ at 700∼800 ms; 132 Hz and 3985 $\mu A/cm^2$ at 1000∼1100 ms.

data and generalize across states. Experience replay allows for more stable policy updates and can be beneficial when dealing with diverse scenarios. In TD3, the implementation of exploration strategies, such as noise injection in the action space, also proves to be effective in handling various initial states.

While the software-driven approach used in this research provides significant convenience as a controlled testing environment, it may face challenges when applied to in vivo models due to the complexity of biological systems and the variability of individual responses. The discrepancy in performance between the BGT network and real-world electrophysiological signals could stem from several factors such as data distribution differences, noise and artifacts, sampling rate mismatch, and feature variability. To alleviate such differences, the feature extraction methods have been preliminarily verified in real signals. Future research should address these limitations by considering additional brain nuclei in the pathological network to overcome the discrepancies between simulated and actual in vivo conditions as well as reveal other potential numerical features that contribute to RL training, e.g., gamma and theta band oscillations [38]. Utilizing a personalized and electrophysiological-based neural simulation model, as suggested in [39], might also facilitate more effective customization of parameter adjustments to individual differences. For the RL agent model, domain adaptation techniques such as transfer learning or adversarial training can bridge the gap between training and deployment domains. Further, fine-tuning the model on small datasets from individual patients also enhances its adaptation to their specific characteristics. We will assess the performance of the model through preclinical animal experiments and refine our research with the aforementioned methods in the future.

Following our in silico simulations, future research will focus on integrating the RL model into a cl-DBS system. This involves embedding the model in low-latency systems, such as FPGAs or embedded processors, to enable real-time data processing and feedback loops, and validating the system through preclinical animal studies. To fit the RL model within the constraints of battery-operated or embedded devices, techniques such as model compression, pruning, and quantization will be employed. Effective integration will also require incorporating real-time monitoring and safety mechanisms to prevent overstimulation and ensure continuous adaptability. Pre-deployment testing and simulation using neural simulators will help validate the model's performance under in vivo conditions. Additionally, cloud-based processing and strict adherence to regulatory standards will ensure robust and safe operation. Longitudinal studies will be essential for adapting the model to evolving patient conditions. These combined efforts will bridge the gap between simulation and real-world application, paving the way for advanced DBS treatments.

V. Conclusion

This study presents a significant advancement in the application of cl-DBS for Parkinson's patients. By instantiating a basal ganglia-thalamic (BGT) model and designing it as an interactive RL-friendly environment, we established four finely tuned RL agents (SAC, TD3, PPO, A2C) for comprehensive comparison.

The major findings highlight the remarkable efficacy of the optimized TD3 architecture, which demonstrated a substantial 67% reduction in average power dissipation compared to the open-loop system. Notably, this reduction was achieved while preserving the normal response of the BGT network, showcasing the potential for improved energy efficiency in cl-DBS. TD3 effectively mitigated thalamic error responses under pathological conditions and exhibited optimal performance to achieve complete power savings under healthy conditions. These results underscore the significance of our adaptive parameter tuning for optimizing therapeutic effects.

The integration of RL algorithms into DBS controllers represents a promising avenue for advancing neuromodulation therapies. These controllers offer dynamic and adaptable parameter tuning, enhancing the precision and efficacy of stimulation. The envisioned future development and deployment of such controllers hold the potential to revolutionize DBS treatments, offering personalized and optimized interventions tailored to individual patient needs.

REFERENCES

[1] T. Lebouvier et al., "The second brain and Parkinson's disease," *Eur. J. Neurosci.*, vol. 30, no. 5, pp. 735–741, Sep. 2009.

[2] *Parkinson's Disease Foundation*. Accessed: Jan. 2024. [Online]. Available: https://www.parkinson.org/Understanding-Parkinsons/Statistics

[3] W. Dauer and S. Przedborski, "Parkinson's disease: Mechanisms and models," *Neuron*, vol. 39, no. 6, pp. 889–909, Sep. 2003.

[4] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol. Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[5] W. Xu, G. S. Russo, T. Hashimoto, J. Zhang, and J. L. Vitek, "Subthalamic nucleus stimulation modulates thalamic neuronal activity," *J. Neurosci.*, vol. 28, no. 46, pp. 11916–11924, Nov. 2008.

[6] M. Parastarfeizabadi and A. Z. Kouzani, "Advances in closed-loop deep brain stimulation devices," *J. NeuroEngineering Rehabil.*, vol. 14, no. 1, pp. 1–20, Aug. 2017.

[7] F. Alonso-Frech et al., "Non-motor adverse effects avoided by directional stimulation in Parkinson's disease: A case report," *Frontiers Neurol.*, vol. 12, Jan. 2022, Art. no. 1756286419838096.

[8] V. Gómez-Orozco, I. De La Pava Panche, A. M. Álvarez-Meza, M. A. Álvarez-López, and Á. Á. Orozco-Gutiérrez, "A machine learning approach to support deep brain stimulation programming," *Revista Facultad Ingeniería Universidad Antioquia*, vol. 89, no. 95, pp. 20–33, Dec. 2019.

[9] M. Lu, X. Wei, Y. Che, J. Wang, and K. A. Loparo, "Application of reinforcement learning to deep brain stimulation in a computational model of Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 339–349, Jan. 2020.

[10] D. Krylov, R. Tachet des Combes, R. Laroche, M. Rosenblum, and D. V. Dylov, "Reinforcement learning framework for deep brain stimulation study," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Yokohama, Japan, Jul. 2020, pp. 2847–2854.

[11] Q. Gao et al., "Model-based design of closed loop deep brain stimulation controller using reinforcement learning," in *Proc. ACM/IEEE 11th Int. Conf. Cyber-Phys. Syst. (ICCPS)*, Apr. 2020, pp. 108–118.

[12] H. Agarwal and H. Rathore, "Novel reinforcement learning algorithm for suppressing synchronization in closed loop deep brain stimulators," in *Proc. 11th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2023, pp. 1–5.

[13] G. Brockman et al., "OpenAI gym," 2016, *arXiv:1606.01540*.

[14] R. S. Sutton et al., *Reinforcement Learning: An Introduction*, 2nd ed., Cambridge, MA, USA: MIT Press, 2018.

[15] J. E. Rubin and D. Terman, "High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model," *J. Comput. Neurosci.*, vol. 16, no. 3, pp. 211–235, May 2004.

[16] D. Terman, J. E. Rubin, A. C. Yew, and C. J. Wilson, "Activity patterns in a model for the subthalamopallidal network of the basal ganglia," *J. Neurosci.*, vol. 22, no. 7, pp. 2963–2976, Apr. 2002.

[17] R. Q. So, A. R. Kent, and W. M. Grill, "Relative contributions of local cell and passing fiber activation and silencing to changes in thalamic fidelity during deep brain stimulation and lesioning: A computational modeling study," *J. Comput. Neurosci.*, vol. 32, no. 3, pp. 499–519, 2012.

[18] M. Rizzone, "Deep brain stimulation of the subthalamic nucleus in Parkinson's disease: Effects of variation in stimulation parameters," *J. Neurol., Neurosurg. Psychiatry*, vol. 71, no. 2, pp. 215–219, Aug. 2001.

[19] R. Ramasubbu, S. Lang, and Z. H. T. Kiss, "Dosing of electrical parameters in deep brain stimulation (DBS) for intractable depression: A review of clinical studies," *Frontiers Psychiatry*, vol. 9, p. 302, Jul. 2018.

[20] A. D. Dorval, A. M. Kuncel, M. J. Birdno, D. A. Turner, and W. M. Grill, "Deep brain stimulation alleviates parkinsonian bradykinesia by regularizing pallidal activity," *J. Neurophysiol.*, vol. 104, no. 2, pp. 911–921, Aug. 2010.

[21] P. Gorzelic, S. J. Schiff, and A. Sinha, "Model-based rational feedback controller design for closed-loop deep brain stimulation of Parkinson's disease," *J. Neural Eng.*, vol. 10, no. 2, Apr. 2013, Art. no. 026016.

[22] *Gymnasium Documentation*. Accessed: Jul. 2024. [Online]. Available: https://gymnasium.farama.org/

[23] M. S. Okun, "Deep-brain for parkinson's disease," *New England J. Med.*, vol. 367, no. 16, pp. 1529–1538, Oct. 2012.

[24] T. M. Herrington, J. J. Cheng, and E. N. Eskandar, "Mechanisms of deep brain stimulation," *J. Neurophysiol.*, vol. 115, no. 1, pp. 19–38, 2016.

[25] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, Sep. 1970.

[26] S.-B. Lee et al., "Predicting Parkinson's disease using gradient boosting decision tree models with electroencephalography signals," *Parkinsonism Rel. Disorders*, vol. 95, pp. 77–85, Feb. 2022.

[27] S. Little and P. Brown, "What brain signals are suitable for feedback control of deep brain stimulation in Parkinson's disease?" *Ann. New York Acad. Sci.*, vol. 1265, no. 1, pp. 9–24, Aug. 2012.

[28] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 278, no. 6, pp. H2039–H2049, Jun. 2000.

[29] J. E. Fleming and M. M. Lowery, "Changes in neuronal entropy in a network model of the cortico-basal ganglia during deep brain stimulation," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5172–5175.

[30] M. F. Anjum, S. Dasgupta, R. Mudumbai, A. Singh, J. F. Cavanagh, and N. S. Narayanan, "Linear predictive coding distinguishes spectral EEG features of Parkinson's disease," *Parkinsonism Rel. Disorders*, vol. 79, pp. 79–85, Oct. 2020.

[31] T. Haarnoja et al., "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. ICML*, Stockholm, Sweden, 2018, pp. 1861–1870.

[32] S. Fujimoto et al., "Addressing function approximation error in actor-critic methods," in *Proc. ICML*, Stockholm, Sweden, 2018, pp. 1587–1596.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[34] V. Mnih, "Asynchronous methods for deep reinforcement learning," in *Proc. ICML*, New York, NY, USA, 2016, pp. 1928–1937.

[35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[36] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. ICLR*, San Juan, Puerto Rico, May 2016.

[37] J. Schulman et al., "Trust region policy optimization," in *Proc. ICML*, Lille, France, 2015, pp. 1889–1897.

[38] E. M. Adam, E. N. Brown, N. Kopell, and M. M. McCarthy, "Deep brain stimulation in the subthalamic nucleus for Parkinson's disease can restore dynamics of striatal networks," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 19, May 2022, Art. no. e2120808119.

[39] C. M. Davidson, A. M. de Paor, H. Cagnan, and M. M. Lowery, "Analysis of oscillatory neural activity in series network models of Parkinson's disease during deep brain stimulation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 86–96, Jan. 2016.