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

Chia-Hung Cho[1], Chii-Wann Lin[1], Pin-Jui Huang[1], and Meng-Chao Chen[1]

[1]Affiliation not available

January 29, 2024

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

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

***Abstract*— *Objective:* 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. *Methods:* 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. *Results:* 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. *Significance:* 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).**

I. Introduction

PARKINSON'S disease (PD) is a chronic neurodegenerative disease that impacts the central nervous system. It ranks as the second most prevalent neurodegenerative disease, primarily targeting the motor neuron system, following Alzheimer's disease [1]. Globally, there are over 10 million individuals living with PD [2]. Parkinson's Disease (PD) is characterized by the degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNc) [3]. The reduction in dopamine levels due to the degeneration results in primary motor symptoms, such as tremors, limb rigidity, bradykinesia, and postural instability [4]. Non-postural symptoms include mood changes, depression, and other emotional alterations, as well as challenges related to swallowing, chewing, speaking, and urinary and skin problems. Levodopa/L-dopa medications are effective in the early stages, but their efficacy diminishes as the disease progresses, leading to motor complications. At this critical juncture, deep brain stimulation (DBS) becomes a promising advanced treatment option. High-frequency ($\geq$100 Hz) DBS has proven effective in regulating the activities of stimulated downstream nuclei [5].

DBS can be broadly classified into two categories: open-loop (ol-DBS) and closed-loop (cl-DBS). The significant distinction lies in the parameter tuning approach. In the open-loop system, commonly used in clinical practice, physicians manually adjust parameters. In contrast, cl-DBS utilizes discriminative biomarkers, enabling automatic parameter regulation through tailored algorithms. While the open-loop system is characterized by a simpler architecture and lower resource requirements, it has drawbacks such as increased time and power consumption, subject dependency, and a reduced battery lifespan [6]. Among these concerns, overstimulation of deep brain regions presents serious concerns, leading to adverse effects like dystonia, dyskinesia, freezing of gait, or pathological laughter [7].

Incorporating machine learning into closed-loop design enables the automatic extraction of patterns and features in brain signals for decision-making, eliminating the need for manually crafted predefined signal processing rules. This approach is beneficial for various aspects, including computer-aided diagnosis/detection (CAD) for DBS candidate selection, optimization of cl-DBS algorithms, surgical targeting, etc [8], [9]. Our primary focus lies in optimizing the cl-DBS algorithm, addressing the crucial post-surgical challenge of DBS device programming for therapy [10].

As a subfield of machine learning, reinforcement learning (RL) holds considerable strength in dealing with dataset-free and dynamic decision problems, which sub-consequently preserves all costs in signal acquisition. It enables an agent to learn how to perceive and interpret an interactive environment and take suitable action to maximize reward through trial and error. Such a process is akin to biological learning systems. Despite significant research and applications in fields such as robotics, gaming, and autonomous driving, the utilization of RL in the medical domain has yet to be thoroughly explored.

Our study aims to develop a cl-DBS algorithm for PD using RL to enhance DBS parameter adjustment, as depicted in

Chia-Hung Cho is with the Department of Biomedical Engineering, National Taiwan University, Taipei 100, Taiwan (R.O.C.) (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 (R.O.C.) (email: i309505013.eic09g@nctu.edu.tw).

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

*Chii-Wann Lin is with the Department of Biomedical Engineering, National Taiwan University, Taipei 100, Taiwan (R.O.C.), Biomedical Technology and Device Laboratories, Industrial Technology Research Institute, Hsinchu 310, Taiwan (R.O.C.) as a senior research consultant, and also with Center for Artificial Intelligence Research, University of Tsukuba as a visiting professor. (correspondence email: cwlinx@ntu.edu.tw).

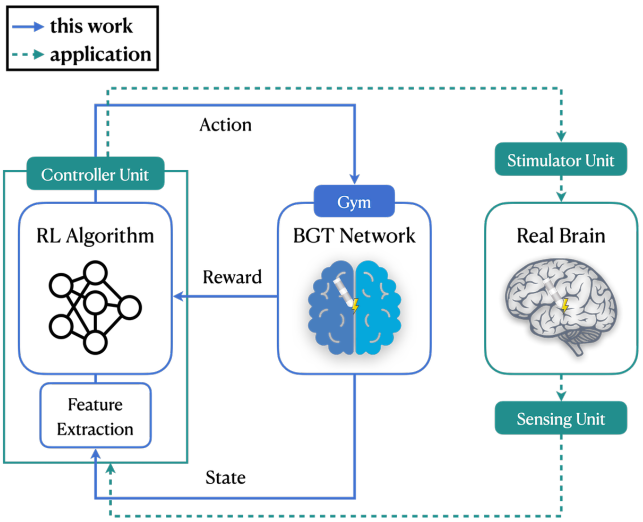

Fig. 1. The overall architecture of this study. The solid blue lines represent what we will implement in this work, whereas the dashed green lines represent a practical direction for the future. Both present the closed-loop characteristics.

Figure 1. Initially, we construct an interactive Basal Ganglia-Thalamic (BGT) network environment based on the Rubin-Terman model [11]–[13], simulating brain dynamics in both healthy and pathological states. Subsequently, representative biomarker signals are identified for feature extraction, enabling the training of an adaptive agent to tune parameters through interaction with the environment. A comparative analysis is conducted among four mainstream RL training frameworks—Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (TD3), Proximal Policy Optimization (PPO), and Advantage Actor-Critic (A2C)—to observe distinctive characteristics and potential in the context of DBS parameter adjustment. Results demonstrate the effectiveness and superiority of our TD3-based method in terms of power efficiency and error response.

II. RL-based CL-DBS Related Works

This section outlines recent research utilizing reinforcement learning techniques for PD treatment via DBS.

Lu et al. [14] established a basal ganglia network from So et al.'s work [13] as the RL exploration environment. Additionally, they integrated a Cerebellar Model Articulation Controller (CMAC) neural network into the actor-critic RL framework, offering benefits in nonlinear function approximation. The findings indicate that the RL-based DBS method consumes 63.3% of the energy compared to open-loop DBS, demonstrating the capacity to restore distorted relay reliability in the thalamus.

Krylov et al. [15] effectively applied RL-based suppression to regular, chaotic, and bursting collective oscillations modeled using Bonhoeffer–van der Pol oscillators and the Hindmarsh–Rose neuronal model. They employed Proximal Policy Optimization (PPO) to train RL agents, enabling them to learn optimal policies for suppressing synchronous neuronal activity.

Gao et al. [16] introduced a Markov decision process (MDP) model to capture the dynamics of neuron activities in the Basal-Ganglia network. Employing convolutional neural networks (CNNs) within the actor-critic architecture, they extracted features from the time series input. Their results showcased the alleviation of PD symptoms with an average stimulation frequency of 45 Hz, indicated by a decrease in the error index and spectral density within the beta band.

Agarwal et al. [17] introduced an RL algorithm, leveraging TD3, to suppress synchronization in neuronal activity during episodes of neurological disorders with reduced power consumption. The proposed framework undergoes a comparative analysis against other RL algorithms, specifically the A2C, the Actor-critic with Kronecker-featured trust region (ACKTR), and the Proximal Policy Optimization (PPO). This evaluation is conducted on an ensemble of oscillators, including the Bonhoeffer-van der Pol and Hindmarsh-Rose models.

Acknowledging the inherent limitations of neurobehavioral simulations in fully capturing real brain pathology, we offer a conductance-based (Hodgkin-Hoxley) environment for comparative analysis of different RL frameworks, encompassing both on-policy and off-policy gradient methods [18]. Furthermore, most studies have overlooked the potential coexistence of pathological and healthy states which may result in a scenario where the RL agent exclusively identifies the disease state, leading to overstimulation. Additionally, 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). We prioritize the utilization of well-established and validated features (refer to Section III-C). This ensures that these features retain their effectiveness when applied to real-world signals.

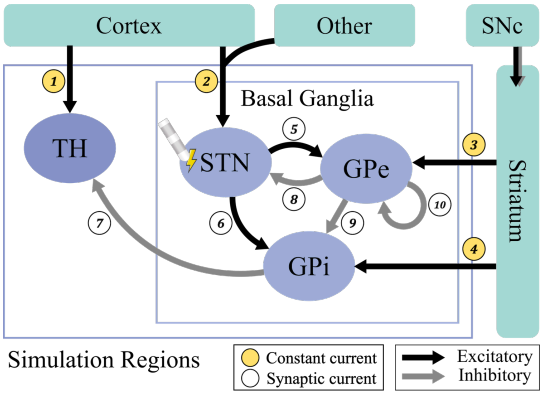

Fig. 2. Illustration of the simulated regions (purple box) and the related currents. Purple ovals are the four neuron types in the basal ganglia-thalamus (BGT) network, containing 10 neurons in each nucleus. Excitatory inputs are represented by black arrows, including ① input from the sensorimotor cortex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

III. METHODS

A. BGT Network Model Simulation

Firstly, we establish the "BGT Network," illustrated in Figure 1, 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 Figure 2.) Each nucleus comprises 10 neurons to balance fidelity and computational efficiency. The parameters and equations of this biophysics model are originated from the work by So et al. [13] and are implemented in Python. For a comprehensive understanding of equations and parameters, please consult the supplementary material. The BGT network simulation encompasses both normal/healthy and PD/pathological conditions.

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:

$$\begin{aligned} C_m \frac{dv_{STN}}{dt} =& -I_L - I_{Na} - I_K - I_T - I_{Ca} - I_{AHP} \\ & - I_{GPe\to STN} + I_{app(STN)} + I_{DBS}, \end{aligned} \tag{1}$$

$$\begin{aligned} C_m \frac{dv_{GPe/i}}{dt} =& -I_L - I_{Na} - I_K - I_T - I_{Ca} - I_{AHP} \\ & - I_{STN\to GPe/i} - I_{GPe\to GPe/i} + I_{app(GPe/i)}, \end{aligned} \tag{2}$$

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

In the neuronal models, the term $C_m dv/dt$ represents the capacitive current responsible for charging the membrane capacitance C_m in STN, GPi, GPe, and TH-type neurons. The 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 "afterhyperpolarization" 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} (depicted in Figure 2), influence subsequent elements in the model.

We place the I_{DBS} term in (1) since assuming the DBS electrode is placed in the STN region and delivered stimulation waveforms. 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 (cf. figure 3). Maintaining a "charge-balanced" condition 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 diminishes with an increase in pulse width [19]. Additionally, maintaining a fixed pulse width helps minimize charge injection and reduce power consumption [20]. The trained RL agent will determine additional stimulation parameters, such as frequency and amplitude.

TH neurons do not exist intrinsically firing properties in the absence of sensorimotor inputs, I_{SM}. I_{SM} is modeled as a series of anodal, monophasic current pulses with an amplitude

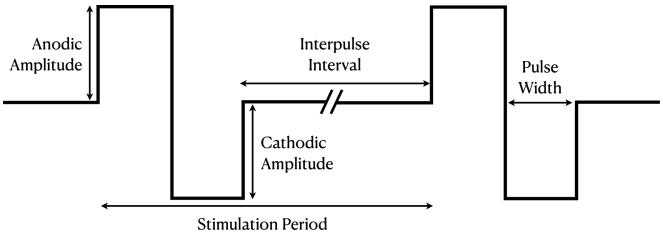

Fig. 3. Illustration of the biphasic, charge-balanced, symmetric DBS pulses we applied throughout our simulation work.

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 should respond faithfully, and promptly to the periodic input with a single action potential (AP) [13]. This signal will subsequently be transmitted to the brainstem and spinal cord to facilitate the execution of actions. In essence, relay error exhibits a high correlation with motor symptoms, as indicated in [21]. It functions as a quantitative metric for assessing PD severity in our study. We quantify the degree of response error using the Error Index (EI), which is formalized as:

$$EI = \frac{N_{error}}{N_{SM}}. \tag{4}$$

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 higher PD severity 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}). Each STN neuron receives inhibitory input from two GPe neurons. Additionally, each GPe or GPi neuron receives excitatory input from two STN neurons and inhibitory input from two other GPe neurons. The effect of the overall BG network and external DBS is propagated to TH through GPi, i.e., $I_{GPi\to TH}$, providing us to evaluate the efficacy of stimulation through the quantified EI.

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. In the literature [11]–[13], they decreased the amount of I_{app} in the PD state for the outcome of fitting physiological signals, and so does our work. Based on the PD etiology, reducing the I_{app} level elucidates the effect of insufficient dopamine secretion by SNc since currents from other brain regions or striatum are correspondingly lessened.

B. Biomarker Selection

In the BGT network, we call for a discriminative signal as the environmental output. Given the varied performance of TH in different states, we hypothesize that the $I_{GPi\to TH}$ synaptic currents carry distinct signal representations. $I_{GPi\to TH}$ is comprised of: $I_{GPi\to TH} =$

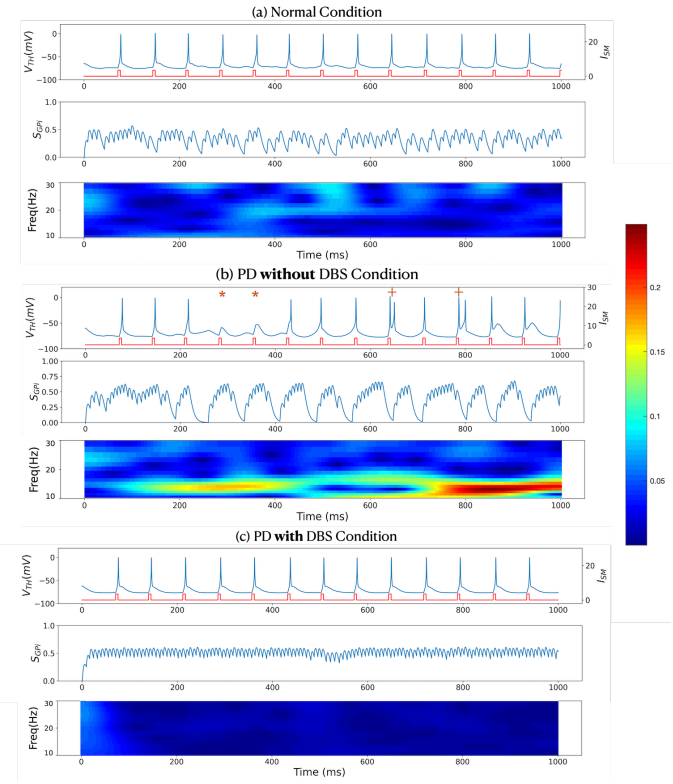

Fig. 4. Thalamus voltage traces, synaptic input signals from GPi to TH ($\boldsymbol{S_{GPi}}$), and scalogram within the beta band in three conditions: (a) normal/healthy (EI=0.0), (b) PD without DBS (EI=0.5), and (c) PD with DBS conditions(EI=0.0). $\boldsymbol{I_{SM}}$ inputs are highlighted in red pulse. +: represents a "bursting" error response (generating more than one AP); *: represents a "missing" error response (TH neuron signal does not constitute an AP). Scalograms are calculated through continuous wavelet transform with the default Morse wavelet. 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).

$g_{GPi\to TH}[v_{TH} - E_{GPi\to TH}]\sum S_{GPi}$, 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. [22], setting S_{GPi} as a biomarker signal. We can further examine the correlation between the S_{GPi} signals and TH membrane potentials in three different states through figure 4. In figure 4 (b), substantial synchronous in GPi nuclei is sufficient to affect thalamic activity through large periodic oscillations/fluctuations in the S_{GPi} signal, which results in a higher EI compared to (a) and (c). The applied DBS can stabilize the S_{GPi} signal, restore faithful relay reliability, and reduce the error response of TH neurons to I_{SM}.

C. Problem Formulation

We customized the OpenAI Gym [23] source, devising a tailored interface that encompasses appropriate action space, state space, reward function, total episode length, and step length. As an initial condition, the environment randomly assigns a state from healthy and PD when an episode starts, aiding in replicating the irregular occurrence of PD.

1) Step Length: The step length significantly influences the time resolution of the action and 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.

2) Action: Action space is composed of the DBS frequency and amplitude value in a total dimension of 2. These values serve as the output of the RL agent and the next state input to the BGT environment. Several studies evaluated the effects of variation in the DBS parameters and suggested suitable ranges ([19], [20], [24], [25]). Borrowing from those works, 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. Initially, we will set the action space to range from -1 to 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. Subsequently, the value will be denormalized back to the desired range within the BGT environment.

3) State: The state space comprises the feature extraction value extracted from the biomarker signal S_{GPi} (as detailed in section III-B) in a total dimension of 6. Contrary to the action space, it serves as the input of the RL model and the output of the BGT environment. We adopt the following extracellular-based feature extraction techniques.

- Signal standard deviation.
- Hjorth Parameters: Hjorth Parameters, comprising activity, mobility, and complexity, offer efficient statistical characterization of time-domain signals. Initially developed for EEG analysis due to low computational complexity [26], they have proven effective in enhancing PD diagnosis with an accuracy of up to 89.3% [27].
- 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 Parkinson's Disease (PD) patients and motor impairments such as bradykinesia/rigidity [28]. PD patients exhibit elevated beta power spectra in both STN and GPi neurons, but these can be suppressed by adequate stimulation amplitude or medication. Figure 4illustrates the substantial difference in the scalograms of S_{GPi} under distinct conditions.
- Sample Entropy (SampEn): SampEn has proven effective in evaluating the complexity of physiological time-series signals and diagnosing disease states [29]. Its advantages over approximate entropy (ApEn), such as data length independence and ease of implementation, make it a preferable choice. A lower entropy value indicates 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 [30].

In figure 5, scaled values of the above feature extraction methods are depicted for both synaptic signal, S_{GPi} from the BGT environment and the EEG dataset from [31] across PD and control (healthy) states. The observed consistency in trends between PD and healthy states suggests promising potential for their application in subsequent agent deployments.

4) Episode Termination Prerequisites: It is inherently logical to conclude an episode when the state is deemed to be "sufficiently optimal," as defined by the following criteria:

- Current EI 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.

5) Reward: Crafting rewards based on reliability components (EI) can be a beneficial approach, considering that reducing thalamus EI is one of the primary objectives in 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:

$$r_1 = EI_{t-1} - EI_t. \tag{5}$$

r_1 could be referred to as a revised score if EI_t dropped and a penalty term if EI_t increased. Next, we structure the energy expenditure penalty using the root mean square of $I_{DBS}(t)$, where the frequency and amplitude components are actions output from the RL model, as:

$$\begin{aligned} r_2 &= -I_{RMS} \\ I_{RMS} &= \sqrt{\frac{1}{T}\int_0^T I_{DBS}^2(t)dt}, \end{aligned} \tag{6}$$

where T denotes the duration of the I_{DBS} stimulation on STN neurons. To expedite the model's adherence to the termination condition without intentionally prolonging the episode, we introduce a "current state penalty" based on the current EI:

$$r_3 = -EI_t \tag{7}$$

Finally, a compensation value for switching off the DBS in healthy states to encourage the model to conserve energy,

$$r_4 = \begin{cases} 1, & \text{if} \quad r_1 \cap r_2 \cap r_3 = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Jointly, the shaped reward function is: $R(t) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3 + \lambda_4 r_4$, where λ are weighting coefficients, in our case, with $\lambda_1 = 250,\ \lambda_2 = 0.01,\ \lambda_3 = 15,\ \lambda_4 = 10$. The weighting coefficients can be tuned based on the importance of each component.

D. RL Actor-Critic Frameworks Implementation

In this study, we evaluate the BGT environment using the Soft Actor-Critic (SAC [32]), Twin Delayed Deep Deterministic Policy Gradient (TD3 [33]), Proximal Policy Optimization (PPO [34]), and Advantage Actor-Critic (A2C [35]) frameworks. All models share the same critic and actor architecture, implemented using PyTorch [36].

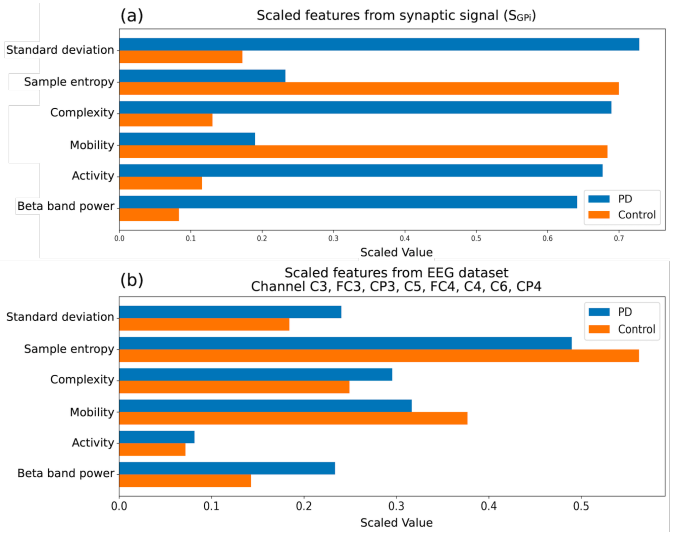

Fig. 5. Preliminary observations on the effect of feature extraction in synaptic signals and EEG signals. In subfigure(b), we select the channels situated above the primary motor cortex (C3, FC3, CP3, C5, FC4, C4, C6, CP4) from the Iowa dataset in [31] for verification. Note that normalization is required for further use of these features.

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 [37]) 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 [38]) 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.

IV. Results

Figure 6 and Figure 7 illustrate the control strategies performed by agents trained using the SAC, TD3, PPO, and A2C RL frameworks in the PD and healthy state. Stimulation 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. The corresponding reduced percentage compared to ol-DBS and average EI for each framework are summarized in Table I.

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. Figure 6 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 on-policy PPO and A2C agents, the resulting actions exhibit heightened variability, and the adjustments in parameters 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 state, guided by the reward design, the agents are expected to minimize or deactivate stimulation to conserve energy without inducing side effects. SAC maintains a stable output with small amplitudes, and the application of stimulation does not result in side effects or an increase in EI. Remarkably, under the control of the TD3 agent, it effectively modulates the amplitude to zero, indicating the cessation of stimulation. This control strategy demonstrates an optimal strategy. 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.

In Table I, the bottom two rows present average EI values for the scenarios without DBS intervention and under ol-DBS for comparison. Especially in the ol-DBS scenario, the elevation of EI in the 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 quantitatively, 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.

TABLE I

Evaluation metrics for all trained RL agents, open-loop DBS (ol-DBS), and without DBS in PD and healthy conditions.

		PD state		Healthy 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

*In ol-DBS regime, $I_{\acute{DBS}}(t)$ is with frequency of 130 Hz and amplitude of 2500 $\mu A/cm^2$. The reduced percentage is calculated by: $1 - (I_{\hat{RMS}}/I_{\acute{RMS}}) \times 100\%$, where $I_{\hat{RMS}}$ is for root mean square of target $I_{DBS}(t)$, while $I_{\acute{RMS}}$ is for the corresponding value in ol-DBS.

V. Discussion

All RL-based cl-DBS algorithms encounter the challenge of establishing an effective interaction environment. We focused on the core dynamic changes of action potentials at the bottom layer of nerve cells in the BGT network. This simulation allows us to capture information between cell synapses, treating them as biomarker signals and extracting features for use as inputs in the RL framework. In contrast to other related literatures, we employ random mechanisms to simulate scenarios where pathological and normal states intermittently emerge in the design of the environment. Additionally, a BGT-Gym framework is provided, allowing for modifications to action and state spaces, reward design, etc. The chosen feature extraction methods, validated in extracellular electrophysiological signals (as seen in literature [26]–[30], etc.), imply their feasibility for future deployment. Considering additional brain nuclei in the pathological network may reveal more numerical features with potential for RL training, including gamma and theta band oscillations [39]. Utilizing a personalized and electrophysiologically-based neural simulation model, such as the one in [40], facilitates more effective customization of parameter adjustments to individual differences.

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

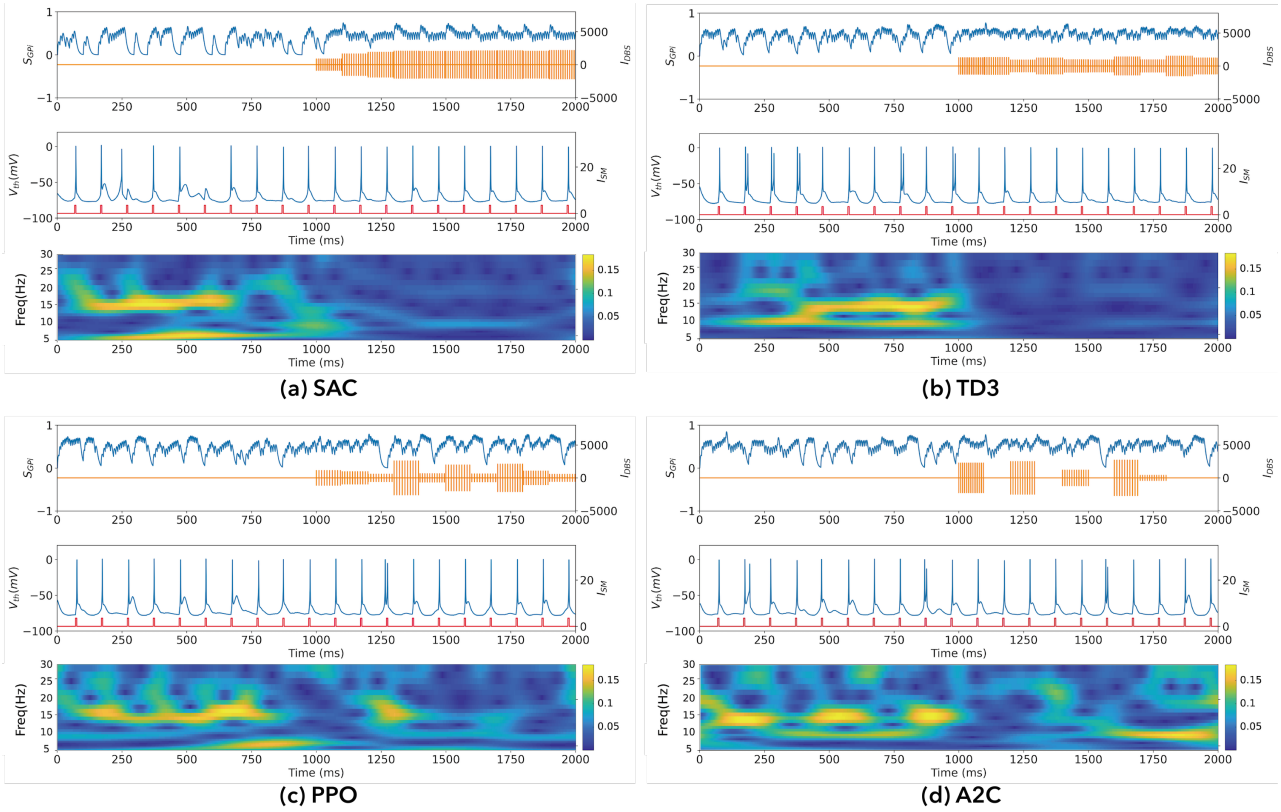

Fig. 6. Control strategy in the **PD** state by (a) SAC, (b) TD3, (c) PPO, and (d) A2C RL agents. Stimulation is activated after 1000 milliseconds. Each subplot includes the biomarker signal ($\boldsymbol{S_{GPi}}$), action signal ($\boldsymbol{I_{DBS}}$), thalamus action potentials, sensorimotor input ($\boldsymbol{I_{SM}}$), and the scalogram of the $\boldsymbol{S_{GPi}}$ signal in the beta frequency band, from top to bottom.

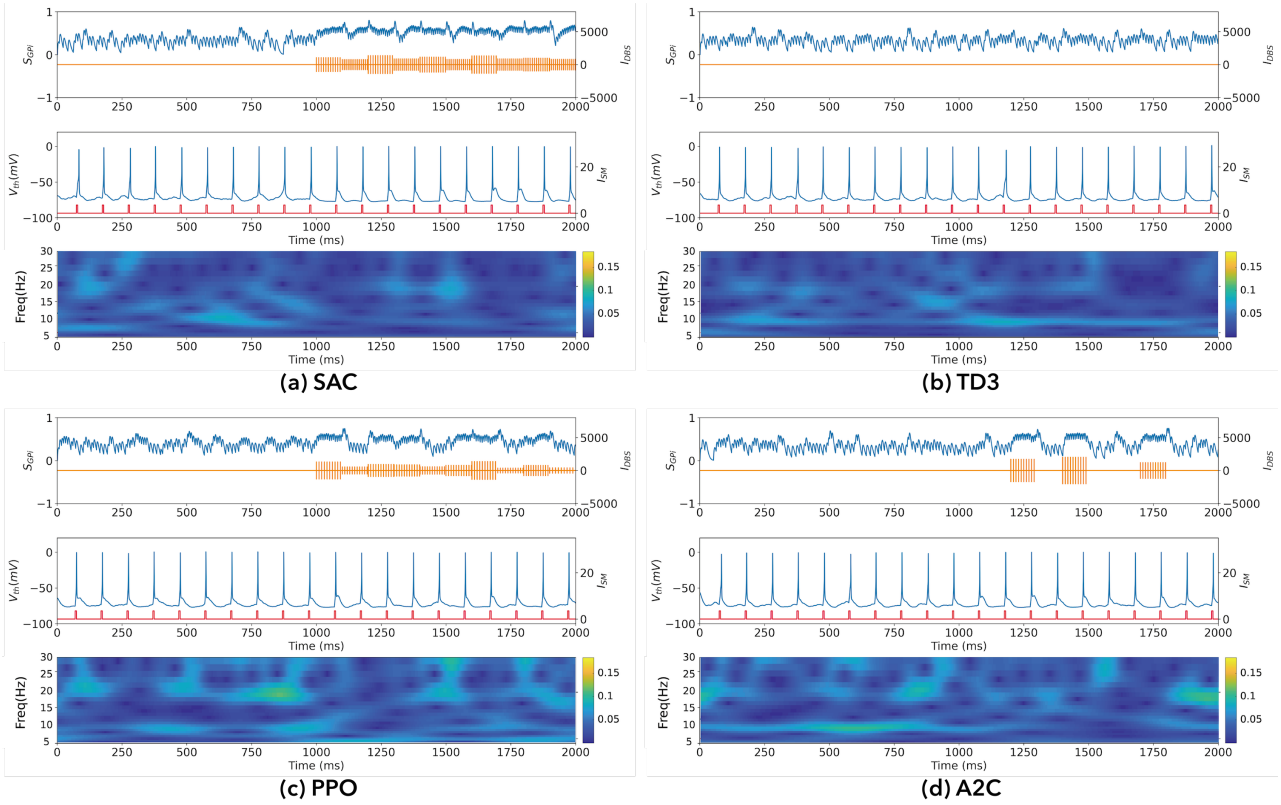

Fig. 7. Control strategy in the **healthy** state by (a) SAC, (b) TD3, (c) PPO, and (d) A2C RL agents. Stimulation is activated after 1000 milliseconds. Each subplot includes the biomarker signal ($\boldsymbol{S_{GPi}}$), action signal ($\boldsymbol{I_{DBS}}$), thalamus action potentials, sensorimotor input ($\boldsymbol{I_{SM}}$), and the scalogram of the $\boldsymbol{S_{GPi}}$ signal in the beta frequency band, from top to bottom.

such as noise injection in the action space, also proves to be effective in handling various initial states.

VI. Conclusion

This study presents a significant advancement in the application of cl-DBS for PD. 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.

Acknowledgment

We gratefully acknowledge the support and resources provided by National Science and Technology Council (Grant: MOST 111-2221-E-002 -079 -MY3).

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, Available at: https://www.parkinson.org/Understanding-Parkinsons/Statistics, Accessed Jan 2024.

[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, Apr 2008.

[5] W. Xu *et al.*, "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 Neuroeng 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," *Front Neurol*, vol. 12, pp. 1756286419838096, Jan 2022.

[8] J. Watts *et al.*, "Machine Learning's Application in Deep Brain Stimulation for Parkinson's Disease: A Review," *Brain Sci*, vol. 10, no. 11, pp. 809, Nov 2020.

[9] W.J. Neumann and M.C Rodriguez-Oroz, "Machine Learning Will Extend the Clinical Utility of Adaptive Deep Brain Stimulation," *Mov Disord*, vol. 36, no. 4, pp. 796–799, Apr 2021.

[10] V. Gómez-Orozco *et al.*, "A machine learning approach to support deep brain stimulation programming," *Rev Fac Ing*, no. 95, pp. 20–33, Apr-Jun 2020

[11] 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-Jun 2004.

[12] D. Terman *et al.*, "Activity patterns in a model for the subthalamopallidal network of the basal ganglia," *J Neurosci*, vol. 22, no. 7, pp. 2963–2976, Apr 2002.

[13] R.Q So *et al.*, "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, Jun 2012.

[14] M. Lu *et al.*, "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.

[15] D. Krylov *et al.*, "Reinforcement Learning Framework for Deep Brain Stimulation Study," *IJCAI-20th*, Yokohama, Japan, pp. 2847–2854, Jul 2020.

[16] Q. Gao *et al.*, "Model-Based Design of Closed Loop Deep Brain Stimulation Controller using Reinforcement Learning," *In 2020 IEEE/ACM 11th ICCPS*, pp.108–118, 2020.

[17] H. Agarwal et al., "Novel Reinforcement Learning Algorithm for Suppressing Synchronization in Closed Loop Deep Brain Stimulators," *11th Int. IEEE/EMBS Conf. Neural Eng (NER)*, pp. 1–5, Apr 2023.

[18] RS Sutton *et al.*, "Reinforcement learning: An introduction," in *The MIT Press*, 2nd ed. Cambridge, MA, 2018.

[19] M. Rizzone *et al.*, "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.

[20] R. Ramasubbu *et al.*, "Dosing of electrical parameters in deep brain stimulation (DBS) for intractable depression: a review of clinical studies,"*Front Psychiatry*, vol. 9, pp. 302, Jul 2018.

[21] A.D Dorval *et al.*, "Deep brain stimulation alleviates parkinsonian bradykinesia by regularizing pallidal activity," *J Neurophysiol*, vol. 104, no. 2, pp. 911-921, Aug 2010.

[22] P. Gorzelic *et al.*, "Model-based rational feedback controller design for closed-loop deep brain stimulation of parkinson's disease," *J Neural Eng*, vol. 10, no. 2, pp. 026016, Apr 2013.

[23] G. Brockman *et al.*, "Openai gym," *arXiv preprint arXiv:1606.01540*, Jun 2016.

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

[25] T.M. Herrington *et al.*, "Mechanisms of deep brain stimulation," *J Neurophysiol*, vol. 115, no. 1, pp. 19–38, Jan 2016.

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

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

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

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

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

[31] MF Anjum *et al.*, "Linear predictive coding distinguishes spectral EEG features of Parkinson's disease," *Parkinsonism Relat Disord*, vol. 79, pp. 79–85, Oct 2020.

[32] T. Haarnoja *et al.*, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *ICML-35th*, Stockholm, Sweden, PMLR 80, 2018.

[33] S. Fujimoto *et al.*, "Addressing function approximation error in actor-critic methods," *ICML-35th*, Stockholm, Sweden, PMLR 80, 2018.

[34] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, Jul 2017.

[35] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," *ICML-33rd*, New York, USA, 2016.

[36] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv preprint arXiv:1912.01703*, Dec 2019.

[37] T.P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *ICLR-4th*, San Juan, Puerto Rico, May 2016.

[38] J. Schulman *et al.*, "Trust Region Policy Optimization," *ICML-31st*, Lille, France, 2015.

[39] E.M. Adam *et al.*, "Deep brain stimulation in the subthalamic nucleus for Parkinson's disease can restore dynamics of striatal networks," *Proc Natl Acad Sci USA*, vol. 119, no. 19, pp. e2120808119, May 2022.

[40] CM Davidson *et al.*, "Analysis of Oscillatory Neural Activity in Series Network Models of Parkinson's Disease During Deep Brain Stimulation," *IEEE TBME*, vol. 63, no. 1, pp. 86–96, Jan 2016.