 ISSN 2362-0412

Oration

Engineers mimicking Reinforcement Learning in the Brain

by
Dr. N.W.J.A.L. Prins

Department of Electrical and Information Engineering, Faculty of Engineering, University of Ruhuna.

Introduction

Reward based learning, 'reinforcement learning' (RL) is found in many day to day activities. Similar models can be found in the brain. The basal ganglia and the cerebellum in the brain were thought to be dedicated to motor control in the early days. However, now it is believed that the cerebellum, the basal ganglia, and the cerebral cortex are specialized for different types of learning; supervised learning, reinforcement learning and unsupervised learning, respectively (Doya, 2000). Engineers for centuries have mimicked nature and in this case, the three types of learning are types that have been used extensively in artificial intelligence (AI). AI relies on the machine being able to learn certain patterns and structures of the data in order to predict output or the action the machine should do. Self-driven cars, thought controlled robots, prosthetics, photo tagging in social media, google targeted advertisements are a few examples.

In early days, engineers modelled the neural networks in the brain; artificial neural networks (ANN). The biological neural network has many redundancies in its connections, is able to map the input to the output and has the ability to adjust itself based on the feedback received through sensory inputs like touch and visual feedback. Similarly, ANN had many interconnections and had a mechanism to receive feedback by comparing the output to a desired signal. This comparison to a desired signal brings about an 'error' which can be used to change the parameters in the ANN. Such error based learning is what is called supervised learning. In unsupervised learning, no such external signal exists and the system purely relies on the input data structure to make decisions. There are many real world examples that cannot rely on such extremes but rather rely on something in-between. In other words we do not have anything to use as an error signal (as in the case of amputees for controlling robotic limbs), neither can we rely solely on the input data structure in real world cases where the data is contaminated with noise. The reward

based learning in the brain motivated engineers to find a way to introduce this reward based learning in to machine learning. This revolutionized the field of machine learning as it was possible now to teach the machine based on reward and punishment.

Brain-machine interfaces (BMI) is one of the applications where medicine and engineering fuse together. This is widely used in paralyzed patients to help them to think of movements and be able to control computer cursors, artificial limbs, robots or their own paralyzed limbs. Many of the BMIs designed through the ages have used supervised learning techniques creating the necessity of a desired signal in order to perfect the algorithm. However, in the case of amputees or paralyzed individuals, there is no physical movement in order to acquire a desired signal. This is where RL can be helpful as it does not need a supervised error signal. If we are able to acquire a reward signal from the brain itself, the system can be truly autonomous. The initial work was done in 2007 as proof of concept and applied to rodents that were able to control a robotic arm (DiGiovanna et al., 2007). In 2010 the research progressed to value (reward and punishment) based decision making (Mahmoudi and Sanchez, 2011).

This paper reviews the research carried out from 2011 to establish a new RL architecture that can use BMIs for paralyzed individuals to be able to control in a truly autonomous manner. In our research we developed an architecture that was able to take the feedback from the brain itself (Pohlmeyer et al., 2012). Next, we were able to show that the this architecture was able to receive signals from two different sources from the brain to be truly an autonomous agent (Mahmoudi et al., 2013). Extracting features from these neural data is also an important step in the BMI process (Prins et al., 2013). In the next paper, we showed how even amidst neural perturbations, this model was able to self-adjust and maintain its performance with animal data (Pohlmeyer et al., 2014). The next challenge we had was to be able to change the architecture in a way to receive a less than perfect feedback. Our research showed the theoretical simulations of being able to include a confidence metric (Prins et al., 2014). We have shown with simulated data that indeed this is possible (Prins et al., 2017b). One of the research published was behavioural training for non-human primates (NHP) to carry out experiments in order to test such models (Prins et al., 2017a). We had already shown proof of concept that the striatum was a good candidate to extract such a reward signal (Geng et al., 2013). All of this gave us the confidence that a system can be implemented where a paralyzed person will someday be able to go home with a system without having a caregiver supervising this system. We performed some BMI clinical trials with electroencephalographic (EEG) data for spinal cord injured (SCI) subjects (Gant et al., 2018). Final research shows how a SCI subject was implanted with two electrocorticography (ECoG) strips and was sent home

with ability to open and close his hand with his own volition (Cajigas et al., 2021). These 10 research publications span across a decade and capture work done by the author for a longer period (and is continuing). This paper discusses the process from the initial phase where we started mimicking RL in the brain and finally sending a paralyzed person home to be able to open and close his hands with his own volition.

Methods

BMI has several components as shown in Figure 1: data acquisition, signal processing, external device/actuator and feedback. Data acquisition block collects the neural data from the subject and pre-amplifies it before sending it to the signal processing block. The signal processing block is the hub of the BMI. It does all the filtering and feature extraction needed for the algorithm to be able to classify the signals or to be able to interpret the subject's intention. Next is the external device or the actuator (robotic arm, computer cursor). The final component is the feedback provided; this can be audio-visual feedback, feedback to the brain or to the algorithm itself. One of the challenges in BMI is to be able to give the feedback to the classification algorithm. In able bodied persons, this signal is taken from their limbs and given as a feedback. Unlike in an able bodied person, a paralyzed person is unable to move their limbs and hence the necessity for a BMI system.

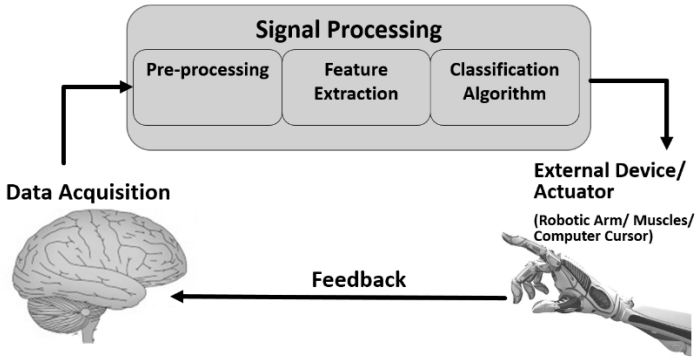

Figure 1: Brain-Machine Interface (BMI) system: Data acquisition, signal processing, external device/actuator and feedback.

Therefore we developed a reward based architecture that would not require such error signals and the preliminary work was done using NHP. Animal behavioural training rely on repetitive movements, encouraged by rewards (usually food pellets or juice rewards) (Prins et al., 2017a). The different types of learning in the brain are shown in Figure 2. Out of these, the supervised learning relies on the external error signal which is difficult to

 ISSN 2362-0412

acquire from paralyzed subjects and the unsupervised learning techniques which rely on the input data structures are not 100% reliable due to noise in biological data. Therefore we proposed the RL based system.

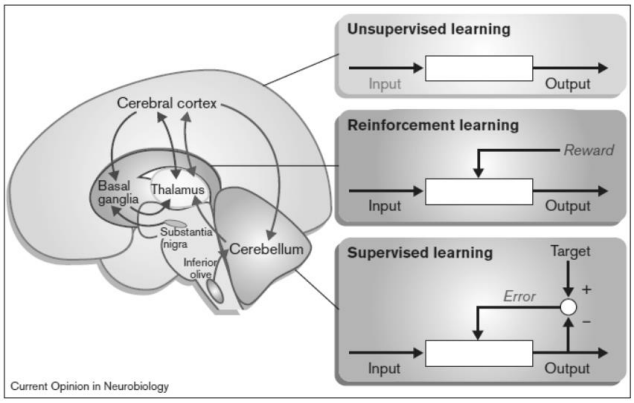

Figure 2: Different types of learning in the brain (Doya, 2000).

RL techniques are vast and can be quite exhaustive; any method one can think of to include a reward in to the algorithm, such method can be used for RL. Therefore the first step was to select a RL architecture which would be a realistic model to test our theory on. We selected the actor-critic RL paradigm where the actor takes an action and the critic criticizes this action (Mahmoudi et al., 2013). Figure 3 shows the basic model for an autonomous system. We were able to show proof of concept using Hebbian RL; neurons that fire together, wire together. Unlike traditional ANN, using the Hebbian rule makes it possible to give the feedback faster to earlier stages of the network (Mahmoudi et al., 2013).

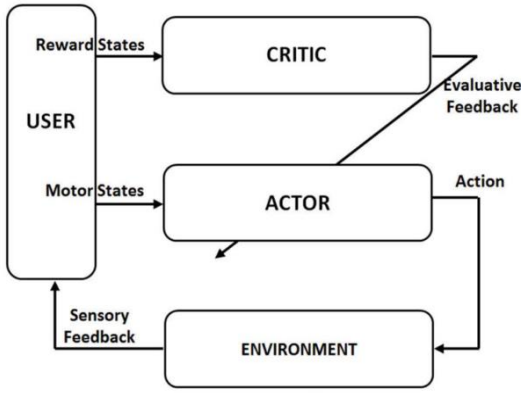

Figure 3: The actor–critic model for autonomous BMI (Prins et al., 2014).

In order to test the stability of this system, we introduced neural perturbations and tested on NHP (Pohlmeyer et al., 2014). Even with a 50% of the neural data missing or 50% neural data addition, the system was able to quickly readjust and maintain its accuracy levels. The critic is the key to making the BMI autonomous; therefore we proposed to take the critic signal from the reward centres of the brain itself. We have shown that this is possible with NHP neural data (Prins et al., 2013). However, one of the discoveries from this research was that the critic accuracy limited the overall performance. Since biological data have less than perfect accuracy, it was important to be able to not rely on the critic completely but only use the critic when it was able to provide good feedback. We overcame this challenge by introducing a confidence metric for the critic (Prins et al., 2014); when the critic was confident, the critic feedback would be used, but otherwise it would be ignored. This system was tested with NHP for a two choice task where the animal was controlling the robotic arm. The experimental paradigm steps are given in Figure 4.

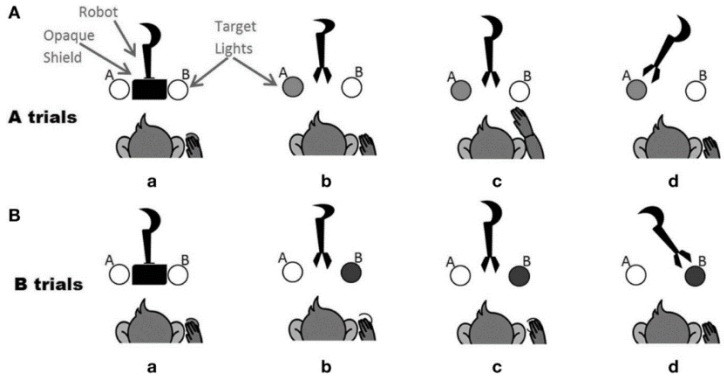

Figure 4: Experimental paradigm for the NHP controlling the robotic arm for two types of trials (Prins et al., 2014).

The next experiments were on simulated data from the reward centres of the brain and testing if the BMI was able to maintain the accuracy with a less than perfect critic (Prins et al., 2017b). Through simulations we were able to find out the sweet spot for the thresholding of the critic confidence (Prins et al., 2017b).

Results

The results presented here are in two sections:(1)foundations of perfect critic and critic with confidence metric for RL based BMI (2)applications of BMI for human clinical trials. Figure 5 shows an example of neural data from NHP. Here the primary motor cortex (MI)

was targeted for the input to the actor and the nucleus accumbens (NAcc) was targeted for the input to the critic. The colour shades show neural data from different neurons.

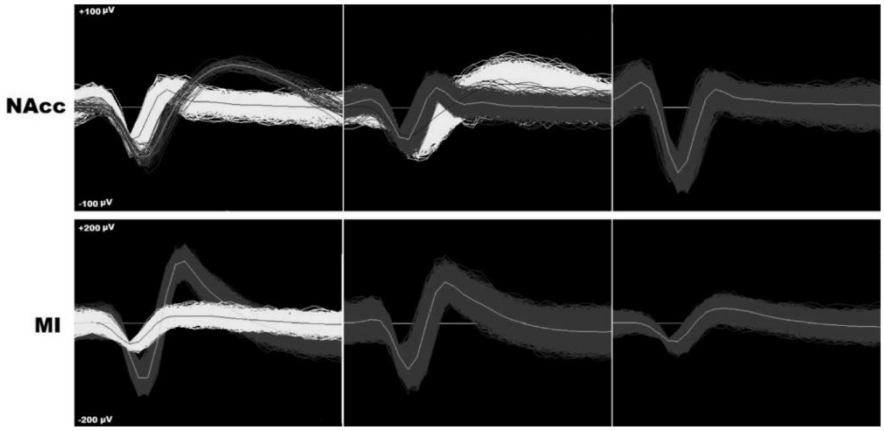

Figure 5: Neural data from reward centre of the brain (NAcc) and motor cortex (MI) (Prins et al., 2017a)

In order for the algorithm to learn the task, there are several parameters that need to adjust. Some of these are called the weights which map the input to the output. Figure 6 shows how the weights were stabilized and when there were perturbations, adjusted itself again to stabilize and give an accurate output. This showed proof of concept that the Hebbian actor-critic model was theoretically suitable for our application. NHP tests with this showed that even with 50% add or drop, the BMI was able to quickly adjust its weights and maintain the performance (Pohlmeyer et al., 2014). Thus proving that for a perfect critic, the actor is able to maintain its performance.

 ISSN 2362-0412

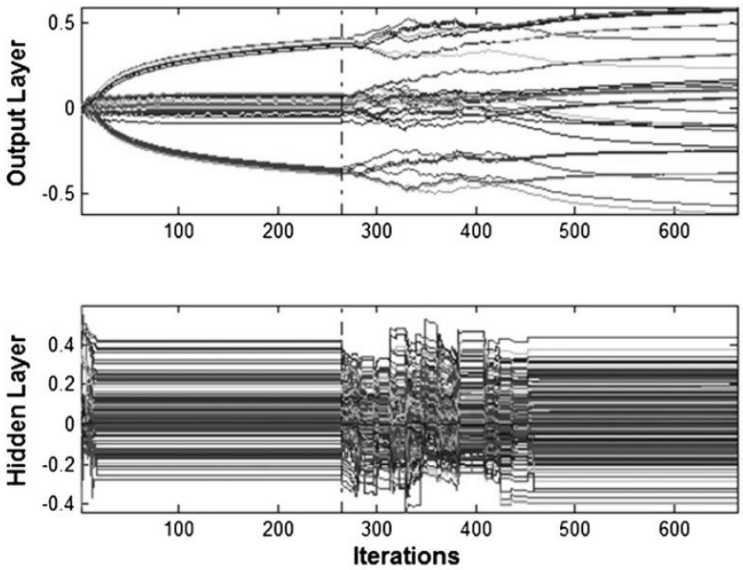

Figure 6: Hidden layer and output layer weights of the ANN for the actor (Mahmoudi et al., 2013)

Next, the critic was tested. Figure 7 shows how having the confidence measure inbuilt to the critic increases the overall performance in (A)simulated data (B)simulated data with noise (C)real NHP data (Prins et al., 2014). This research showed that if we have the ability to include this confidence metric into the critic, the overall performance can be improved.

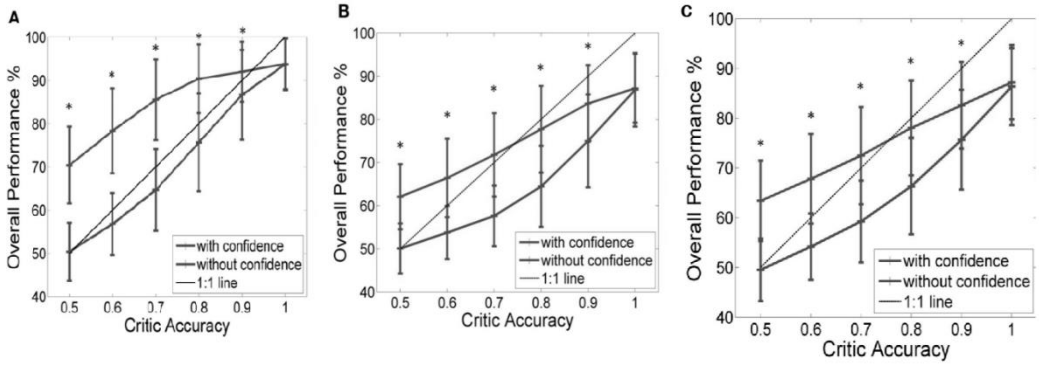

Figure 7: Comparison of overall performance with and without the confidence measure (A) simulated data (B) simulated data with noise (C) real NHP data (Prins et al., 2014)

The solution to this was to introduce a threshold to the critic. Figure 8 shows how the hidden and output weights changed with no threshold, 0.12 and 0.24 threshold. Through this research we were able to conclude that having a threshold is important but a too high threshold affects the overall system negatively. While the threshold is heavily data

 ISSN 2362-0412

dependant, the research findings concluded 0.1-0.15 threshold was best for the data tested (Prins et al., 2017b).

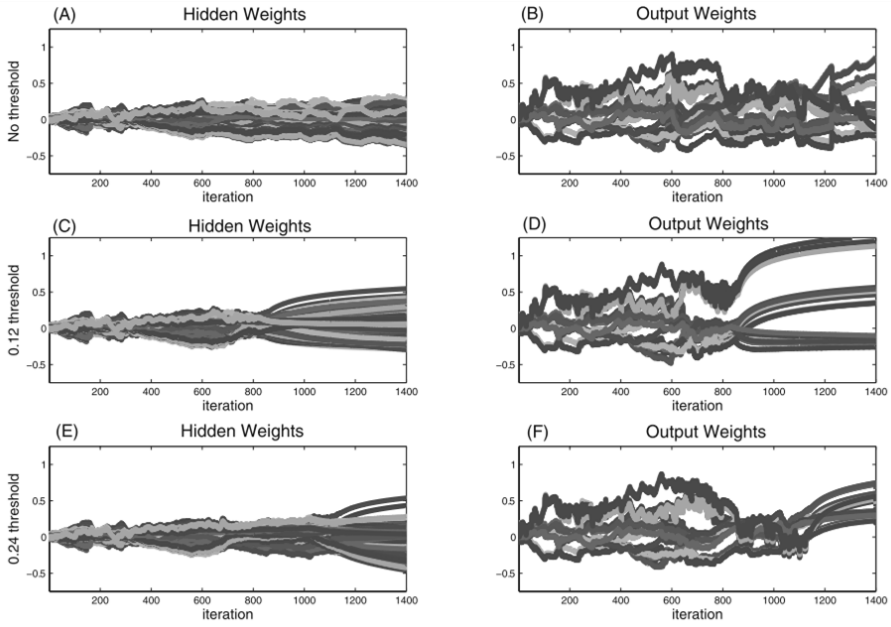

Figure 8: How weights change with the threshold value (Prins et al., 2017b)

After NHP results were confident, we started with non-invasive clinical trials as shown in Figure 9A. Rather than controlling robotic arms (like the NHP did), we were able to give the subjects regain of their own paralyzed hands using functional electrical stimulation (FES) with EEG signals above the motor cortex of C5-C6 SCI subjects. These subjects were able to move their shoulders but had no function in their hands. They were able to control their hand opening and closing with 75% accuracy (Figure 9B) (Gant et al., 2018). While this was impressive, it was not conducive to send subject home with wires connected everywhere.

 ISSN 2362-0412

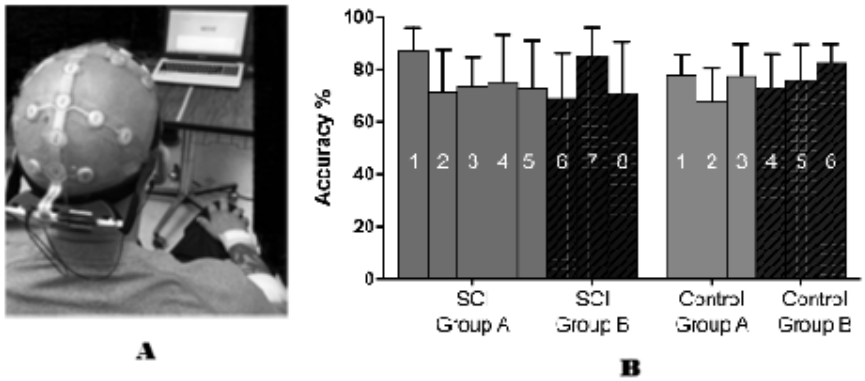

Figure 9: (A) EEG experimental setup showing EEG cap (Data acquisition), signal processing (computer) and actuator (electrode pads on arm) (B) average accuracies of SCI and control groups (Gant et al., 2018).

Final piece of the puzzle was to completely internalize the system and have a person use the system at home (Cajigas et al., 2021). Two ECoG electrode strips were implanted above the motor cortex as decided by fMRI imaging (Figure 10A,B). The signals were transmitted wirelessly from the transmitter. These signals were decoded by the algorithm and fed back to the subject's paralyzed hand by a FES orthosis. Figure 10C shows the experimental setup and Figure 10D shows the spectrograms of the two raw ECoG channels.

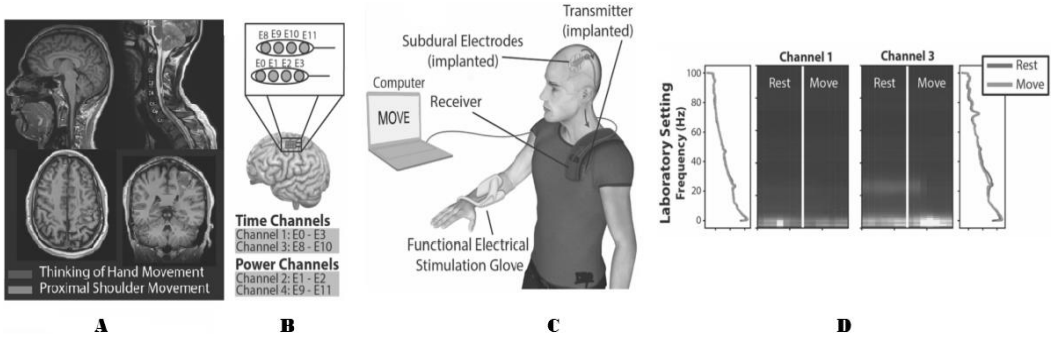

Figure 10: (A) fMRI image showing the active areas during shoulder movement and thinking of hand movement (B) placement of ECoG channels (C) Laboratory set up (D) ECoG spectrograms (Cajigas et al., 2021).

We started at 55% accuracy on a naïve classifier and in the same day achieved 90% accuracy. Over a period of 10 weeks, the BMI maintained an 89% accuracy (Figure 11A). He was able to increase performance and speed of the Jebsen Hand Function Test (JHFT) during the 20 weeks of the study (Figure 11B). His handwriting samples in the figure shows how the fluidity improved also (Figure 11C). Finally, the subject was able to go home with

the system and control it solely with his own volition without an external cue guiding him (Figure 11D).

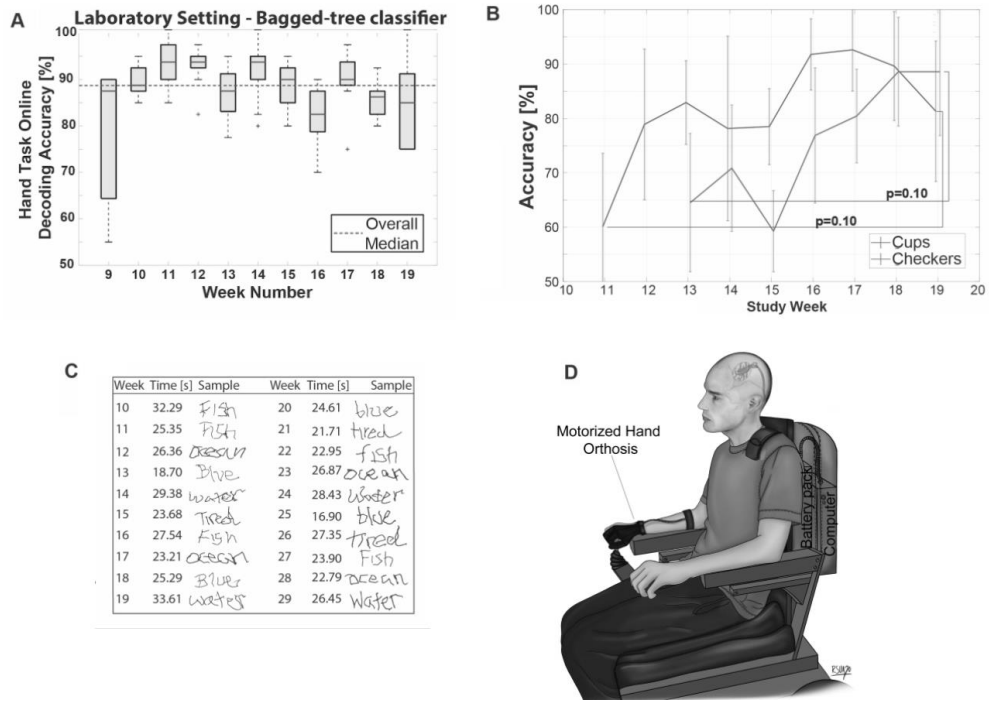

Figure 11: (A) classifier accuracy (B) task accuracy of moving a cup and piece of checkers (C) hand writing (D) Home system (Cajigas et al., 2021).

Conclusion

The research presented here is from theoretical foundations to practical implementation of BMI to be used without an external training signal. Starting with the RL based BMI architecture, it was shown how an actor-critic model is able to handle a reward signal rather than an error from a desired output. This architecture was proven to be effective with input perturbations but sensitive to the critic accuracy. The results showed how a critic confidence measure is able to increase the overall performance. Our results demonstrated that a fully implanted BCI can be safely and reliably used to decode movement intent from motor cortex allowing for volitional control of hand grasp by a patient with SCI in laboratory and home environment. The system was completely internalized in our case where as in many other cases there is a pedestal protruding from the patient's skull.

References

CAJIGAS, I., DAVIS, K. C., MESCHEDE-KRASA, B., PRINS, N. W., GALLO, S., NAEEM, J. A., PALERMO, A., WILSON, A., GUERRA, S. & PARKS, B. A. 2021. Implantable brain-computer interface for neuroprosthetic-enabled volitional hand grasp restoration in spinal cord injury. *Brain Communications*.

DIGIOVANNA, J., MAHMOUDI, B., MITZELFELT, J., SANCHEZ, J. & PRINCIPE, J. Brain-machine interface control via reinforcement learning. Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on, 2007. IEEE, 530-533.

DOYA, K. 2000. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology,* 10**,** 732-739.

GANT, K., GUERRA, S., ZIMMERMAN, L., PARKS, B. A., PRINS, N. W. & PRASAD, A. 2018. EEG-controlled functional electrical stimulation for hand opening and closing in chronic complete cervical spinal cord injury. *Biomedical Physics & Engineering Express,* 4**,** 065005.

GENG, S., PRINS, N. W., POHLMEYER, E. A., PRASAD, A. & SANCHEZ, J. C. Extraction of error related local field potentials from the striatum during environmental perturbations of a robotic arm. 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), 2013. IEEE, 993-996.

MAHMOUDI, B., POHLMEYER, E. A., PRINS, N. W., GENG, S. & SANCHEZ, J. C. 2013. Towards autonomous neuroprosthetic control using Hebbian reinforcement learning. *Journal of neural engineering,* 10**,** 066005.

MAHMOUDI, B. & SANCHEZ, J. C. 2011. A symbiotic brain-machine interface through value-based decision making. *PloS one,* 6**,** e14760.

POHLMEYER, E., MAHMOUDI, B., GENG, S., PRINS, N. & SANCHEZ, J. 2014. Using Reinforcement Learning to Provide Stable Brain-Machine Interface Control. *PLOS One*.

POHLMEYER, E. A., MAHMOUDI, B., GENG, S., PRINS, N. & SANCHEZ, J. C. Brain-machine interface control of a robot arm using actor-critic rainforcement learning. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012. IEEE, 4108-4111.

PRINS, N. W., GENG, S., POHLMEYER, E. A., MAHMOUDI, B. & SANCHEZ, J. C. Feature extraction and unsupervised classification of neural population reward signals for reinforcement based BMI. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013. IEEE, 5250-5253.

PRINS, N. W., POHLMEYER, E. A., DEBNATH, S., MYLAVARAPU, R., GENG, S., SANCHEZ, J. C., ROTHEN, D. & PRASAD, A. 2017a. Common marmoset (Callithrix jacchus) as a primate model for behavioral neuroscience studies. *Journal of neuroscience methods,* 284**,** 35-46.

PRINS, N. W., SANCHEZ, J. C. & PRASAD, A. 2014. A confidence metric for using neurobiological feedback in actor-critic reinforcement learning based brain-machine interfaces. *Frontiers in neuroscience,* 8**,** 111.

PRINS, N. W., SANCHEZ, J. C. & PRASAD, A. 2017b. Feedback for reinforcement learning based brain–machine interfaces using confidence metrics. *Journal of neural engineering,* 14**,** 036016.