SAFETY LEARNING AND ITS UNDERLYING NEURAL MECHANISMS: A DEEP LEARNING APPROACH

By

Monolina Barua Shil

A capstone project submitted for Graduation with University Honors

May 6, 2021

University Honors
University of California, Riverside

APPROVED

Dr. Edward Korzus
Department of Psychology

Dr. Richard Cardullo, Howard H Hays Jr. Chair
University Honors

ABSTRACT

Fear is a fundamental emotion that has a profound influence on how humans interact with and respond to their surroundings. While fear generalization may trigger irrational fear, overgeneralized fear has been correlated with neuropsychiatric vulnerabilities, including Post-Traumatic Stress Disorder (PTSD). Studies have revealed that medial prefrontal cortex (mPFC) circuit-level mechanisms support fear discrimination learning; however, the mechanisms that underlie safety learning to overcome irrational fears remain unclear. Neuroimaging techniques have been integrated with computer analytics to examine the mechanisms of safety learning involving the response to irrational fears after multiple experiences with fearful but not dangerous stimuli. Current methods of extraction of data obtained through large-scale imaging of neural dynamics, using calcium biosensor GCAMP and head-mounted microscopes, are labor-intensive and subject to bias. This project identifies a Deep Learning classification algorithm for artifact removal that allows for accurate disambiguation between Good Neurons and Bad Neurons in mPFC circuit-level mechanisms underlying safety learning. Results highlight the foundation for an efficient procedure for precise data analysis, promoting further research into the neural foundations of fear-related disorders.

TABLE OF CONTENTS

Introduction

Sources of anxiety and inhibitors of success—irrational fears have played a pivotal role in individuals' survival instincts. Fear, in general, is a fundamental feeling that has a profound influence on how humans interact with and respond to their surroundings. When such fears, however, lead the brain to incorrectly label stimuli as dangerous, how does it override irrational feelings and prevent a false reaction? Studies in Behavioral Neuroscience have revealed that medial prefrontal cortex (mPFC) circuit-level mechanisms support fear discrimination learning, allowing individuals to distinguish between safety and danger. The prefrontal cortex is able to gather information from higher cortical associations, sensory regions, and hippocampal, amygdalar, and thalamocortical pathways to form an output signal that is sent out to modulate cortical network ability. These results of these mechanisms can be seen in raw video footage of brain activity of free-moving mice via large-scale imaging of neural dynamics. Data from behavioral testing is quite substantial and must be modified through specific software, CNMFE, to be properly organized for analysis. Though CNMFE is powerful in isolating images of neurons in a network, it often includes artifacts in its final output. These artifacts are cells that are not suitable for data analysis due to insufficient structure and/or performance. Because video files of neural cells are similar in appearance, artifact removal must be conducted manually. The complexity of the method to classify Good and Bad Neurons makes it difficult to develop a procedure that is less labor-intensive, as it sometimes leads to a decline in precision. This study proposes a Deep Learning approach to artifact removal that utilizes machine learning to properly organize data from neural networks.

The overarching goal in research relating to the neural foundations of fear is to investigate the circuit-level mechanisms by which the medial prefrontal cortex supports fear

discrimination learning. *Safety Learning*— the understanding and encoding of behaviors in response to safety and danger into memory—is a developing concept, thus there is a significant lack of research on the topic. The data used to develop the deep learning model, as well as to research this topic, was obtained through animal models. Animal models were conditioned to elicit different responses to stimuli. *Fear Conditioning*, a form of classical conditioning, was the means by which behavior was controlled and modified. This form of learning involves the pairing of an aversive stimulus to either a neutral setting or another independent stimulus. Prolonged exposure to the stimuli ultimately alters the initial behavioral responses. The identification and encoding of a stimulus as fearsome are referred to as *Fear Acquisition*. When conditioned fear responses do not have reinforced exposure to stimuli, the decline in conditioned behavior results in *Fear Extinction*. Successful differentiation between safety and danger that results from conditioned fear responses is indicative of *Fear Discrimination*. Unsuccessful differentiation of such, referred to as *Fear Generalization*, would be the result of conditioned fear responses spreading to similar or related stimuli.

Literature Review

The study of Safety Learning is a combination of a few fundamental phenomena: fear acquisition, fear extinction, and fear conditioning. The fundamental goal is to understand how and where the brain encodes fear. The medial prefrontal cortex is critical in expressing and inhibiting fear behavior; the dorsal region of this area is involved in the expression of fear behavior, while the ventral region is involved in the inhibition of fear behavior. Findings suggest that this area of the brain plays a critical role in encoding fear behavior itself—this area of the brain is able to encode information retrieved from stimuli and crystalize what response was assigned to such stimuli into memory. It is no surprise that the medial prefrontal cortex is

involved in the identification and expression of fear, as this area of the brain is involved in the decision-making and encoding of long-term memories. If this encoding is not accurate there may be several detrimental effects, including the onset of anxiety disorders (Courtin *et al*., 2013). Further studies of the medial prefrontal cortex have revealed that such is very significant in refining the stimuli that handle fear response; specifically, this area of the brain has strong control over the stimuli that are interpreted and encoded. Because of this, the neuronal circuits in the medial prefrontal cortex are essential in ensuring accurate fear behaviors (Korzus, 2015). Inaccurate coding of stimuli often contributes to overgeneralize fear, a clinical condition that is associated with PTSD.

Fear Extinction is a phenomenon that allows individuals to avoid acquiring irrational fears. Studies involving fear extinction have revealed that such is an adaptive form of learning that alters itself in response to the environment; the inability to distinguish between real dangers and paranoia result in anxiety disorders. A review of neuronal circuits (Herry, Cyril, *et al*., 2013) has revealed that fear extinction memories and generalized fear memories may be stored within the same brain areas but involve the support of separate neuronal circuits. Extinction learning involves close relationships between the amygdala, hippocampus, and medial prefrontal cortex. Specifically, the basolateral cortex of the amygdala plays a role in the initial acquisition of fear, but areas of the medial prefrontal cortex are involved in the expression and consolidation of the fear memory. Research in this field has reiterated that neuron circuit pathways and connectivity for fear are found in the basolateral cortex, central amygdala, and medial prefrontal cortex. These areas of the brain are instrumental in fear acquisition and fear expression; despite having roots in the same areas, these two functions, again, involve different neuronal circuits (Tovote, Philip, *et*

al., 2015). This suggests that fear extinction circuits may have the ability to inhibit circuits in fear acquisitions, something that may help reduce fearsome responses to stimuli.

Based on the findings of the aforementioned research, the Korzus Laboratory has utilized *in vivo* calcium imaging of the prefrontal cortex and the hippocampus using head-mounted miniscopes to measure the fluorescent signals of neural activations in the cells of free-moving mice during behavioral training. This yielded substantial data which was then modified and used to train a deep learning network. Training deep learning algorithms to classify and label neuronal cells is particularly difficult, as the images and videos relating to such look extremely similar. Most research conducted on deep learning algorithms and neural networks have been analyzed using clearly identifiable images (i.e. household objects, animals, foods, etc.), thus the development of a model with neurons proved to be rather complex.

Methods

The experimental structure of this project will be divided into two stages: Data Acquisition and Data Analysis. As mentioned earlier, this study utilized animal models to exhibit varying degrees of Safety Learning following conditioning. Mice were placed in a setting with a cage floor that had the capability to release electrical currents to shock them. The shock was administered periodically with either a neutral context or a tone. Prolonged exposure to the stimuli led to the onset of anxiety, a result of fear generalization, in mice that experienced the shock with a tone. After data acquisition, various codes and programs in MATLAB were used to conduct Data Analysis.

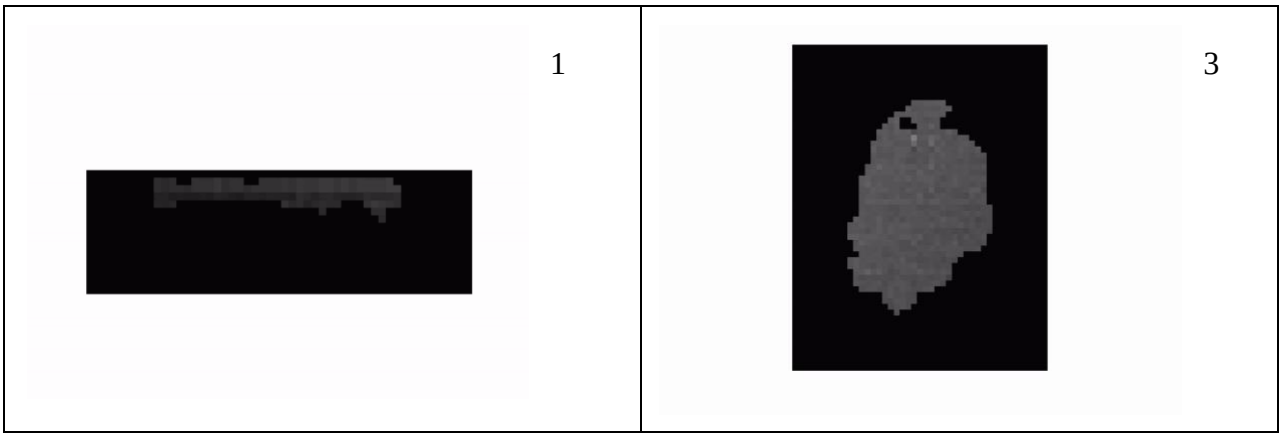
Data Acquisition. Animal models, in this case mice, were subjected to three stages of brain surgery during behavioral testing, allowing for the study of the specific brain areas and neuronal circuits involved in the consolidation of irrational fears. A combination of methods,

optogenetics, electrophysiology, and large-scale imaging of neuronal dynamics, were utilized to perform such tasks. Optogenetics involves the monitoring of neuronal activity through the stimulation of light; protons are sent to cross cell membranes and once a certain threshold has been met, a neuron is fired. This allowed for the study of the excitation or inhibition of specific neurons—analysis of such highlighted the specific neuronal circuits involved in anxiety brought on by irrational fears. Electrophysiology measures the electrical activity of neurons, specifically their action potentials. Like optogenetics, this method provides insight into the specifics of what neuronal circuits are involved in the behavior that is being studied. Lastly, large-scale imaging of neuronal dynamics is time-lapse recordings of neuronal activity in live, free-behaving animal models. This data is the most revealing and is what was utilized in the analysis for this project.

Processing of Raw Data and Data Analysis. The raw footage obtained from the miniscope from behavioral testing highlights the brain activity of mice models in conditioning settings. Calcium imaging through miniscope lenses allows access to neuronal populations in deep regions of the brain of free-moving animals. However, it is computationally challenging to extract single-neuron activity from such data. The video materials obtained from such trials are modified through CNMFE software to extract calcium signals from the data. CNMFE is a code that determines the location and quantity of neurons in a video; after identification, it creates individual neuron files that are analyzed further. It is extremely helpful in demixing and denoising neuronal signals of interest from background fluctuations and neuronal overlaps, efficiently isolating cellular signals. Despite being a powerful software, CNMFE often includes artifacts in its final output. Thus, artifact removal is necessary. Currently, artifact removal is conducted manually through an app on MATLAB; researchers analyze the footprint and trace of a neuron to identify fluorescence and peaking. Such behavior indicates an active cell. . The

existing method for such a procedure is extremely labor-intensive and is subject to substantial bias. This project proposes a computational method that is driven by machine learning. An artificial model of classification eliminates manual labor; the developed model of deep learning is trained to learn classification from preorganized and predetermined sets of videos highlighting both good and bad neurons. The videos utilized in the training are manipulated through a code, such that they are reduced in the number of frames to highlight the maximum fluorescence. Raw footage is converted to twenty-one frames, recognizing only the frames immediately before, during, and immediately after peaking. This ensures that there is not any incorrect labeling of false peaks.

Good Neuron Vs. Bad Neuron. Examples of good and bad neurons are presented below. The main indicator of what should and should not be classified as an artifact is the presence of fluorescence. Simply, fluorescence is the bright flash that is observed in the twenty-one frame video. They can be seen clearly in the gif images 3 and 4 in the figure below. Good Neurons show clear fluorescence, while Bad Neurons (i.e. artifacts) do not. Furthermore, neurons that are irregular in shape and size are labeled as artifacts. In some cases, neurons light up as a result of residual fluorescence from neighboring cells; this also constitutes as an artifact.



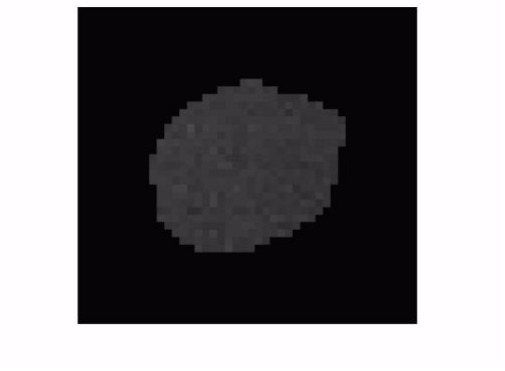
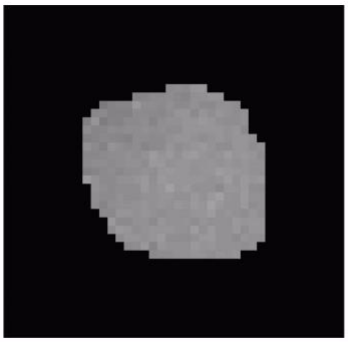
 2	 4

Table 1. *Classification of Neurons.* Image 1 is an example of a Bad Neuron (i.e. an artifact) due to lack of fluorescence; its irregular shape along with its lack of spiking most likely contributes to such. Image 2 is another example of an artifact, as there is no authentic fluorescence, only some that is residual. Images 3 and 4 both represent Good Neurons, as they highlight clear fluorescence in the cell.

Deep Learning. Deep Learning utilizes networks to retrieve useful representations of features directly from data, such that it can be trained to positively identify and classify images. It is a branch of machine learning that is especially suited for image recognition and utilizes neural networks to train and learn from provided material. Different categories of neural networks include Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LTSM). CNNs are effective in areas such as image recognition and classification, as they automatically detect features of images without human supervision and are computationally efficient. LTSMs can process both single data points as well as entire sequences, which can be particularly helpful when it comes to training with video files.

This specific code modified a convolutional network, such that it was suitable for training with video files. This new network created avi files of the raw data and converted such to 21 frames to highlight only points of maximum fluorescence, identified the boundaries around each

cell to isolate individual neurons, found the overlap between them and their surrounding neurons, eliminated those of which had more than 50% overlap (i.e. neurons laying in front of or behind other neurons), and then organized them into their designated categories (i.e. Good Neurons and Bad Neurons). After classification, the frames of data were converted to feature vectors to be prepared for training. Upon completion of the conversion, the model prepared its LTSM network, in which specific parameters were chosen for maximum performance. Parameters of the algorithm included miniBatchsize (i.e. the amount of data included in each sub-epoch weight change; typically, small batch sizes achieve the best generalization performance), learning rate, and validation patience. The miniBatchsize was set to a value of 16, the learning rate was set to 1 e-4, and the validation patience was set to 5. This signifies that a batch size of 16 was utilized to divide the training data set into subsequent groups to calculate and update (i.e. reduce) model errors; a learning rate less than the value of one controlled how quickly the model adapted to the aforementioned errors (a value too low could lead to time-consuming training, while a value too high could lead to suboptimal results); validation patience, which represents the number of times a loss in the validation set can be larger than or equal to the previously smallest loss before the end of network training, was set to 5 and highlighted the number of times this proposed algorithm considered losses in the validation set.

After this setup, the network was trained. Upon completion, the algorithm assembled the video classification network, which added convolutional layers, sequence input layers, and additional LTSM layers. Completion of the assembly of such completed the entire training process and was set for testing.

MATLAB Code Set-Up. The Deep Learning code was assembled specifically for video classifications of good and bad neurons in free-moving animal models that underwent behavioral

conditioning. First, videos were converted to sequences using a pre-trained convolutional neural network, known as GoogLeNet, to extract features from each frame. Pre-trained image classification networks have been previously trained to extract powerful and informative features of data from images. After the conversion of videos, a LTSM network is trained on the sequences to predict labels. Lastly, a network is assembled to classify videos directly by combining layers from both of the aforementioned networks. The general structure of the code is as follows: load training data, define layers, define training options, train, load testing data, determine accuracy. For this project, the deep learning model was trained on four trials. Each trial utilized varying numbers of neurons to identify, which would provide the most accurate results. Trial 1 was trained using 377 neurons (Bad Neurons: 187, Good Neurons: 190), Trial 2 was trained using 300 neurons (Bad Neurons: 150, Good Neurons: 150), Trial 3 was trained using 150 neurons (Bad Neurons: 75, Good Neurons: 75), and Trial 4 was trained using 50 neurons (Bad Neurons: 25, Good Neurons: 25). After training, the models were tested using data that they had not been previously training with. 56 neurons (Bad Neurons: 28, Good Neurons: 28) were tested during each trial for accuracy.

Results

Trial 1 training resulted in a model with 73.68% accuracy. Trial 2 training resulted in a model with 66.67% accuracy. Trial 3 training resulted in a model with 80.00% accuracy. Trial 4 training resulted in a model with 100.00%. Despite varying degrees of accuracy, the models performed about the same during testing. Each model was tested with 56 neurons (Bad Neurons: 28, Good Neurons: 28) it had not been exposed to before for performance. Trial 1 resulted in an overall accuracy of 64.3% in labeling Good and Bad Neurons. Trial 2 resulted in an overall accuracy of 69.7% in labeling Good and Bad Neurons. Trial 3 resulted in an overall accuracy of

59.0% in labeling Good and Bad Neurons. Lastly, Trial 4 resulted in an overall accuracy of 71.5% in labeling Good and Bad Neurons. In Trial 2, Good Neurons were labeled substantially more accurately than Bad Neurons, as labeling of Good Neurons yielded an accuracy of 92.9%, while labeling of Bad Neurons yielded an accuracy of only 46.4%. In contrast, in Trial 4, accuracy in the labeling of Bad Neurons outperformed that in the labeling of Good Neurons; Bad Neurons were classified with 71.5% accuracy, while Good Neurons were classified with 64.3%. Trials 1 and 3 had similar performance results between Good and Bad Neurons, as each yield between about 54%-64% accuracy.

	# of Training Neurons **(Total + # of Bad/Good)**	**Accuracy of Model (%)**	**Testing: Accuracy Overall (%)**	**Testing: Accuracy of Bad Neuron Labels**	**Testing: Accuracy of Good Neuron Labels**
Trial 1	377 (Bad: 187, Good: 190)	73.68%	64.3%	64.3% (18/28)	64.3% (18/28)
Trial 2	300 (Bad: 150, Good: 150)	66.67%	69.7%	46.4% (13/28)	92.9% (26/28)
Trial 3	150 (Bad: 75, Good: 75)	80.00%	59.0%	64.3% (18/28)	53.6% (15/28)
Trial 4	50 (Bad: 25, Good: 25)	100.00%	71.5%	78.6% (22/28)	64.3% (18/28)

Table 2. *Performance of Models Across Trials.* Data analysis reveals that despite an increase in performance during training, testing accuracy remained similar across trials. This suggests alteration of the video files being organized in the algorithm.

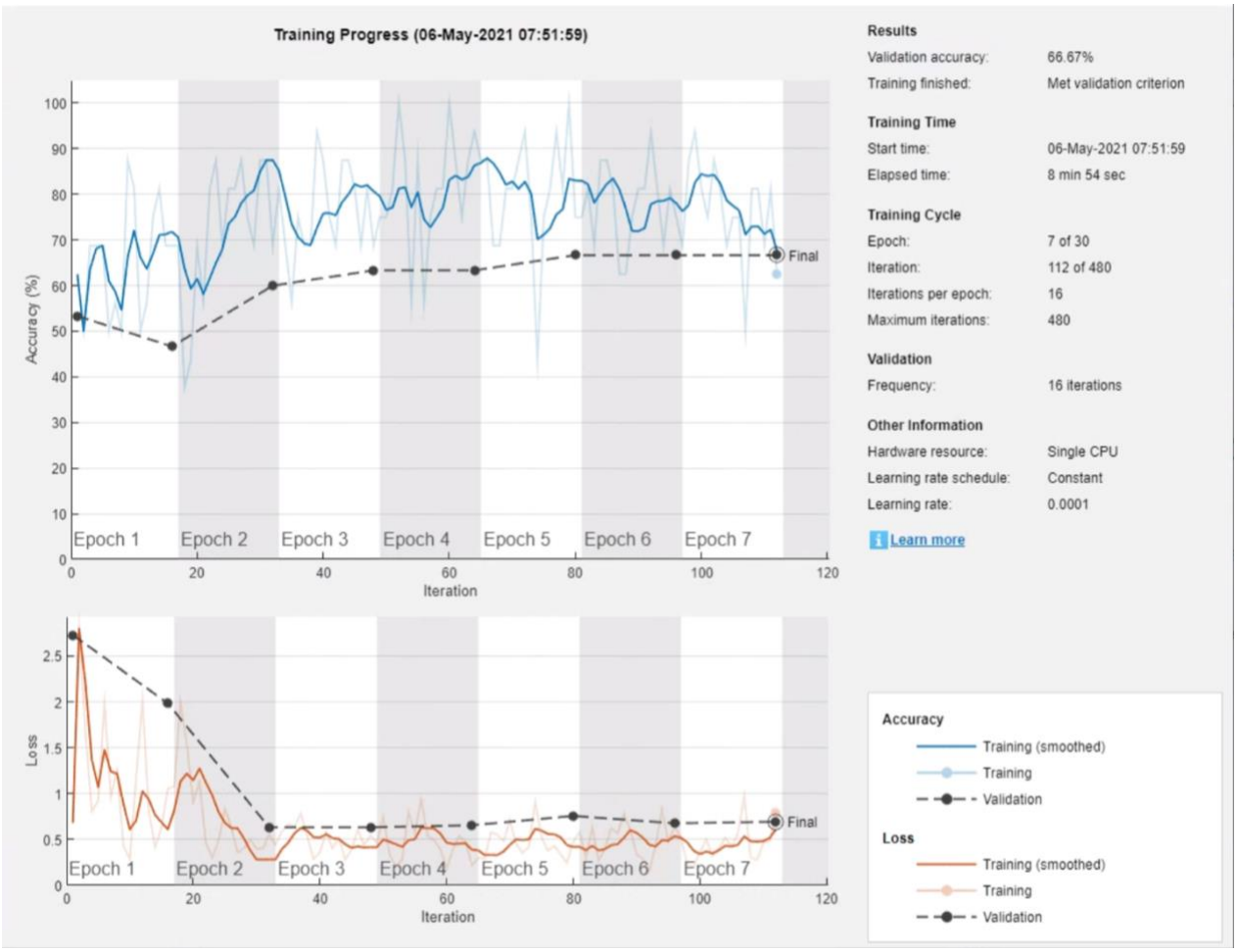


Figure 1. *Deep Learning Training for Trial 1.* Data highlights that training with 377 Neurons, 187 Bad Neurons combined with 190 Good Neurons, yielded an accuracy of 73.68%. This emphasizes that neurons tested on this model should result with the described precision.

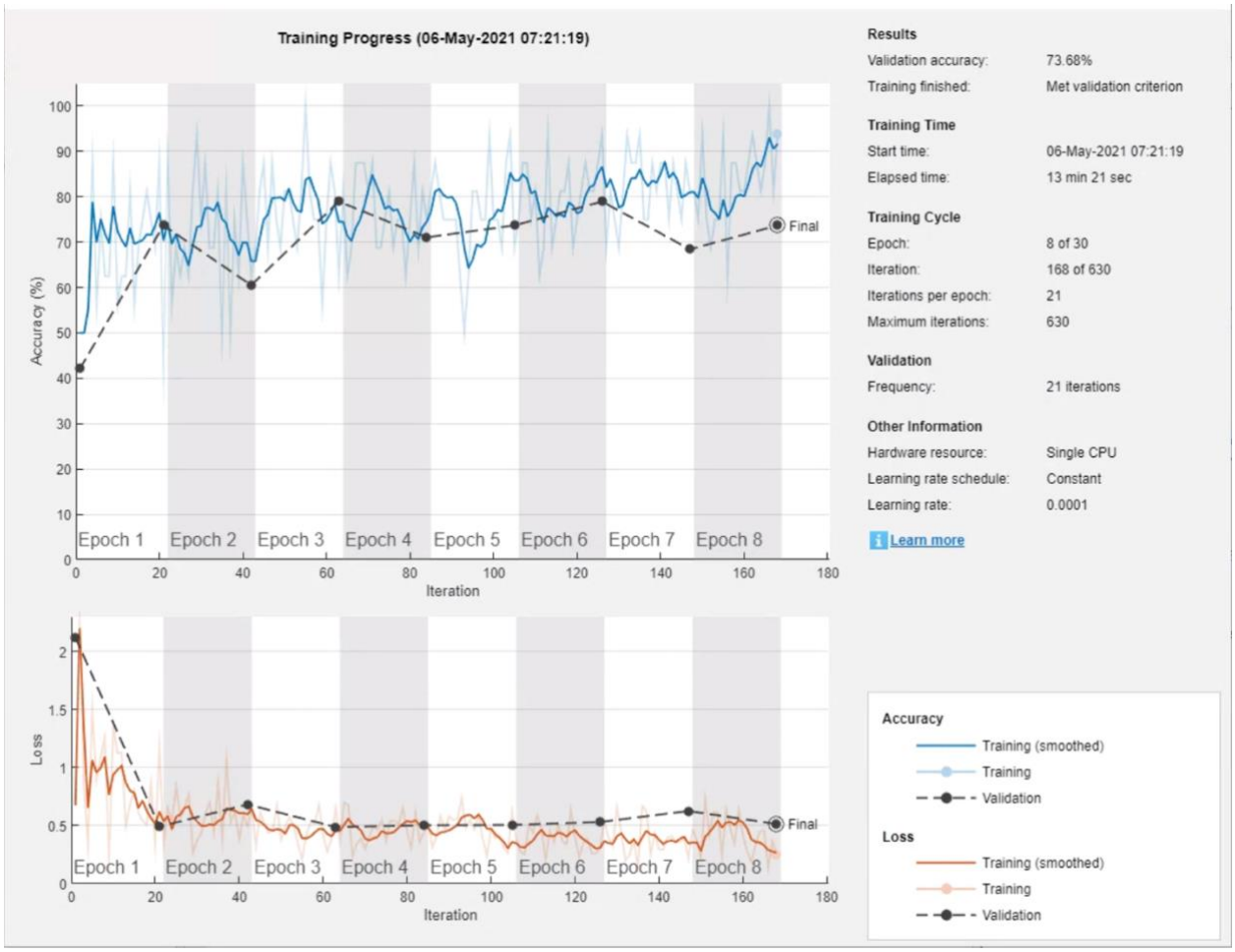


Figure 2. *Deep Learning Training for Trial 2.* . Data highlights that training with 300 Neurons, 150 Bad Neurons combined with 150 Good Neurons, yielded an accuracy of 66.67%. This emphasizes that neurons tested on this model should result with the described precision.

Figure 3. *Deep Learning Training for Trial 3.* Data highlights that training with 150 Neurons, 75 Bad Neurons combined with 75 Good Neurons, yielded an accuracy of 80.00%. This emphasizes that neurons tested on this model should result with the described precision.

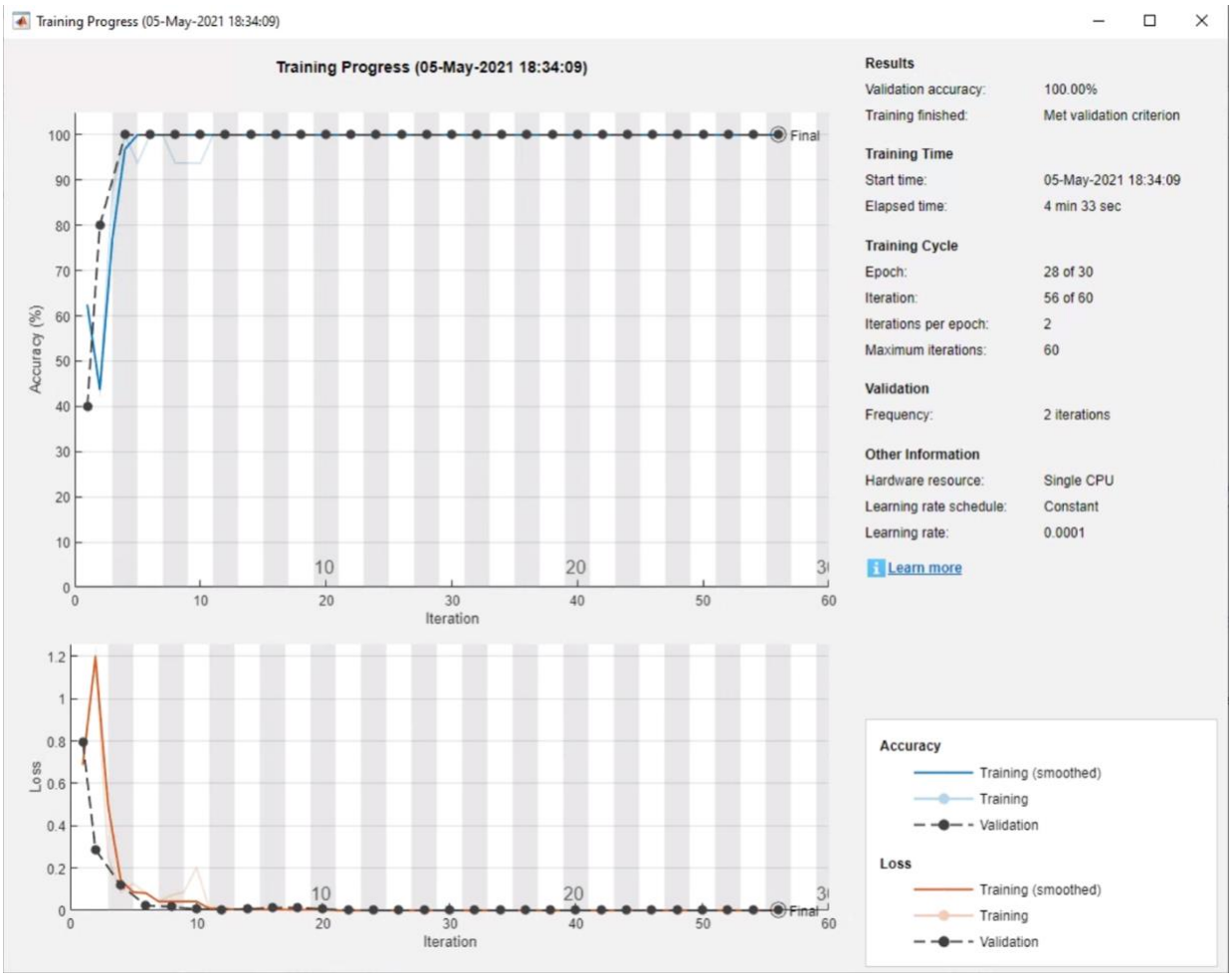


Figure 4. *Deep Learning Training for Trial 4.* Data highlights that training with 50 Neurons, 25 Bad Neurons combined with 25 Good Neurons, yielded an accuracy of 100.00%. This emphasizes that neurons tested on this model should result with the described precision.

Discussion

The primary goal of this study was to identify a mode of artifact removal through a deep learning approach to efficiently classify and organize Good and Bad neurons in a neural network underlying Safety Learning mechanisms. Findings indicate that it is, indeed, possible to conduct analysis through deep learning algorithms. However, it is necessary to alter the structure of the model to yield more consistent, accurate results. The accuracy of the models increased as the number of training neurons decreased, which is unexpected in the development of a Deep Learning Network; theoretically, networks should become more accurate as it is exposed to more data. Though it is not clear why the model is performing in such a fashion, larger quantities of neurons may exhibit more variation possibly making it more difficult for the model to detect a pattern. Furthermore, the training diagrams (i.e., Figures 1-4) reveal that the loss function is not improving over the iterations suggesting that the model may be overfitting or underfitting the data. Essentially, the model is learning the training data but is unable to generalize it to new data. This proposes an alteration to the model parameters to eliminate these discrepancies.

Because there were discrepancies between the projected accuracy of the models and actual performance, future research should alter video processing to create files that are more suitable for the network. One possible method would be to reduce or increase the number of frames of the videos. This may allow for more accurate identification of fluorescence, ultimately leading to precise classification and labeling. Fewer frames may be more helpful, as such would reduce any likelihood of confusion from false spiking and residual fluorescence. Though more frames may lead to more issues, both trials should be tested to retrieve the performance standards of each frame alteration. Alongside this modification, it may be helpful to adjust the parameter settings. Alterations to the miniBatchsize and the learning rate may contribute to more accurate

results in the future. Furthermore, it may be productive to reduce the number of pixels in the neural images to improve both efficiency and accuracy. The deep learning model developed in this study only focused on neurons with 50% overlap, thus it will be helpful to conduct future research into developing a system that isolates all overlapping neurons and labels them accurately. This would require the development of a code that isolates both individual neuron footprint and spiking data from overlapping cells to then be used for training purposes.

Overall, a deep learning network is a complex algorithm that can shift modes of artifact removal in Behavioral Neuroscience research. An accurate developed model can organize and classify neural cells without labor-intensive restraints or biased influences.

REFERENCES

"Classify Videos Using Deep Learning." *MATLAB & Simulink*, www.mathworks.com/help/deeplearning/ug/classify-videos-using-deep-learning.html#ClassifyVideosUsingDeepLearningExample-5.

Courtin, J., et al. "Medial Prefrontal Cortex Neuronal Circuits in Fear Behavior." *Neuroscience*, vol. 240, 2013, pp. 219–242.

"Deep Learning in MATLAB." *Deep Learning in MATLAB - MATLAB & Simulink*, www.mathworks.com/help/deeplearning/ug/deep-learning-in-matlab.html.

Herry, Cyril, and Joshua P Johansen. "Encoding of Fear Learning and Memory in Distributed Neuronal Circuits." *Nature Neuroscience*, vol. 17, no. 12, 2014, pp. 1644–1654.

Herry, Cyril, et al. "Neuronal Circuits of Fear Extinction." *European Journal of Neuroscience*, vol. 31, no. 4, 2010, pp. 599–612.

Korzus, Edward. "Prefrontal Cortex in Learning to Overcome Generalized Fear." *Journal of Experimental Neuroscience*, vol. 9, 2015

Tovote, Philip, et al. "Neuronal Circuits for Fear and Anxiety." *Nature Reviews Neuroscience*, vol. 16, no. 6, 2015, pp. 317–331.