FY20 Naval Innovative Science and Engineering (NISE)/Section 219 Final Report

Fatigue Detection/Prediction using Machine Learning and Wearable Technology

Alex Kniffin, NSWCDD V13

29 January 2021

EXECUTIVE SUMMARY

Fatigue is a known contributor to open water accidents, decreased operational efficiency, and poor Warfighter health. Real-time feedback of the Warfighter's cognitive state will allow for increased awareness of capabilities/limitations and adaptable decision making based on Warfighter readiness. The *Fatigue Detection/Prediction using Machine Learning (ML) and Wearable Technology* project aimed to develop a ML algorithm capable of detecting changes in the Parasympathetic Nervous System (PNS) that are indicative of cognitive fatigue using a Commercial Off-The-Shelf (COTS) wrist-worn device. A biometric dataset of 30 participants (including some active duty personnel) performing quantifiable vigilance tasking was collected and annotated with operator performance metrics and cognitive load. Variations of the Mackworth clock, a vigilance task widely used in psychometric studies to quantify cognitive engagement and fatigue, was used to generate quantitative operator performance metrics and discrete cognitive load states. ML models were trained and validated on the annotated biometric dataset to: 1) regress operator task performance accuracies, and 2) classify cognitive load/task difficulty. A trained Convolution Neural Network (CNN) regression model was able to predict Mackworth Clock task performance accuracy to within a mean absolute error of 2.5%. Additionally, a separate CNN classifier model achieved binary task-type classification accuracies of 86.5%, with different type tasks corresponding to a higher vs. lower cognitive load. The next phase of this Research & Development (R&D) effort will include additional testing events with Navy-relevant tasking (i.e., ship navigation, track management, and other watch standing tasks) with a participant pool of only active duty personnel. The end goal of this effort is to provide a wearable device with accompanying software that is capable of detecting and predicting cognitive fatigue for various Navy-relevant tasking, with the purpose of optimizing Warfighter performance to minimize user error or maximize performance.

OBJECTIVE

Fatigue is very common among Warfighters across all branches of the military. In general, fatigue comes from being chronically overloaded/overworked, deprived of consistent sleep/recovery, and performing monotonous tasking. Warfighters must perform at extremely high levels at all times; however, the conditions in which they are performing are sub-optimal. It is universally understood that fatigue (e.g., cognitive, physical, emotional, etc.) can lead to decreased level of performance, such as poor decision-making and delayed reaction time, both of which are incredibly important parts of the Warfighters job. Unfortunately, military culture does not support the mitigation of fatigue, instead there is an imbedded pride about who has slept the least and asking for a break or more sleep is rarely, if ever, done. Instead, many Warfighters overdose on caffeine and tell peers and supervisors they are fine when they are critically sleep deprived and at a high risk of poor decision-making. This is not an effective mitigation strategy for the chronically fatigued Warfighter and has led to avoidable mistakes. The Warfighter is the most important sub-system of any combat system; therefore, his/her performance should be optimized just as any other sub-system of the combat systems. The Warfighter (a human) is involved in all parts of the military and therefore optimizing his/her performance provides an endless avenue for military applications. Being able to detect and respond to dangerous fatigue levels before a problem occurs would be one possible solution to optimizing performance, which would ultimately save the military millions of dollars and most importantly increase the safety of our Warfighters. The objective of this project is develop a ML algorithm to predict an operator's performance using only biometric data collected from a non-invasive wrist-worn device.

There are multiple tests that are proven to induce cognitive fatigue, as well as, the validation performance metrics collected during these tasks being indicative of cognitive state. These tests are often simple tasks that measure attention, reaction time, and decision-making which can be used as markers for cognitive fatigue or performance decrement. In general, performance results in a given time window that are lower than the individual's "normal" are indicative of cognitive fatigue. These tests are validated with Electroencephalographs (EEGs), which measure brain activity. In the lab environment, EEGs are the current gold standard for measuring cognitive conditions and states such as fatigue; however, they are cumbersome and not ideal for naval environments. Wrist-worn sensors or biomonitors are small form-factor devices that are easily used and may provide useful insight of the Warfighter's physical and cognitive state for leadership awareness and improved strategic decision making. Additionally, the data can be used to justify needed changes or used to monitor Warfighter (individual and team) readiness for system optimization. This includes building a biometric dataset of fatigued participants to train the ML model, which requires designing an Institutional Review Board (IRB) approved protocol for Human Subjects Research to collect data on participants performing tasking known to induce cognitive fatigue.

APPROACH

An unambiguous quantifiable metric of overall general cognitive fatigue is rather difficult to define. One common approach in the literature has been to define cognitive fatigue in terms of a measurable decline in task performance over prolonged effort, or following sustained and acute mental/physical exertion (6, 2). Decrements in tasks

involving sustained attention have also been used to quantify cognitive fatigue (20). This approach has since been validated by studies that have identified a strong correlation between self-identified subjective reporting of cognitive fatigue and the accuracy and processing time for executive tasks (8), and specifically including executive attention tasks requiring individual to resolve conflicting visual information (9).

Additionally, correlations have been found between executive attention, alertness and vigilance tasks on the one hand, and certain physiological signals on the other. Correlations and relationships between EEG activity and auditory alertness tasks (10, 14, 15), as well as, visual alertness (16, 17) have been established, and real time EEG based alertness detection has been demonstrated (1). More recently, a growing body of both and empirical data and neuro-physiological models (24) suggest a correlation between alertness/vigilance tasks and other physiological signals including heart rate variability, and Respiratory Sinus Arrhythmia (RSA) as measured by an Electrocardiogram (ECG) and respiratory recordings (7,19). There is also research to suggest other physiological signal, such as the Galvanic Skin Response (GSR), maybe related to decrements in vigilance tasks (21).

This apparent connection between certain physiological signals, such as heart rate and GSR with alertness, suggests that it may be possible to autonomously detect levels of fatigue, as expressed by a quantifiable decrement in a vigilance task involving executive attention, from purely non-neural physiological signals. Similar detection and recognition of alertness levels from EEG signals have leveraged a variety of Machine Learning techniques including artificial neural networks (22, 23) and Support Vector Machines (SVM) (12). Other research has used more modern Deep Learning architectures such as Constitutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), such as Long Short Term Memory (LSTM) models, to develop EEG based emotion detectors and classifiers (3, 4, 5). This study attempts to expand on this body of work by developing, training, and testing Deep Learning models to detect decrements in vigilance task performance based purely on physiological signal input, and additionally to classify the difficulty of the vigilance task and inferred cognitive load on the participant.

Vigilance Task Accuracy Regression

The sustained executive attention vigilance task chosen for the purposes of this study was the Mackworth clock test. The Mackworth clock was first introduced in 1948 (13) to simulate radar monitoring operations and measure the vigilance decrement of operators as a function of time on task. Initially implemented as a mechanical device, it has since been adapted to a computer based simulation (11), and used widely to measure vigilance and sustained attention task accuracies (18). It involves a dot moving in a circle at regular increments (Figure 1), and at random times undergoing an irregular and discontinuous jumps. The study participants are instructed to acknowledge every time a discontinuous jump has occurred, while simultaneously performing other unrelated tasks. The accuracy with which they correctly acknowledge a discontinuous jump is said to measure a decrement in vigilance.

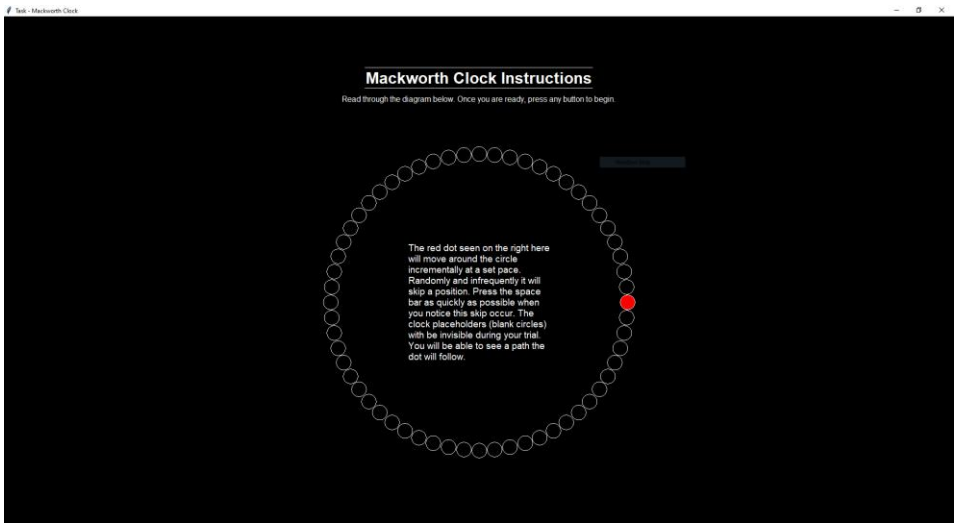

Figure 1. Mackworth Clock Task Screen

The physiological signals used included heart rate, temperature, Blood Pressure Volume (BVP), GSR, and Heart Rate Variability (HRV) as measured by the InterBeat Interval (IBI). The sample rates for the various sample rates were fixed by the manufacturer and were as follows:

1. Heart rate and IBI at 1 Hertz (Hz)
2. Temperature and GSR at 4 Hz
3. BPV at 64 Hz

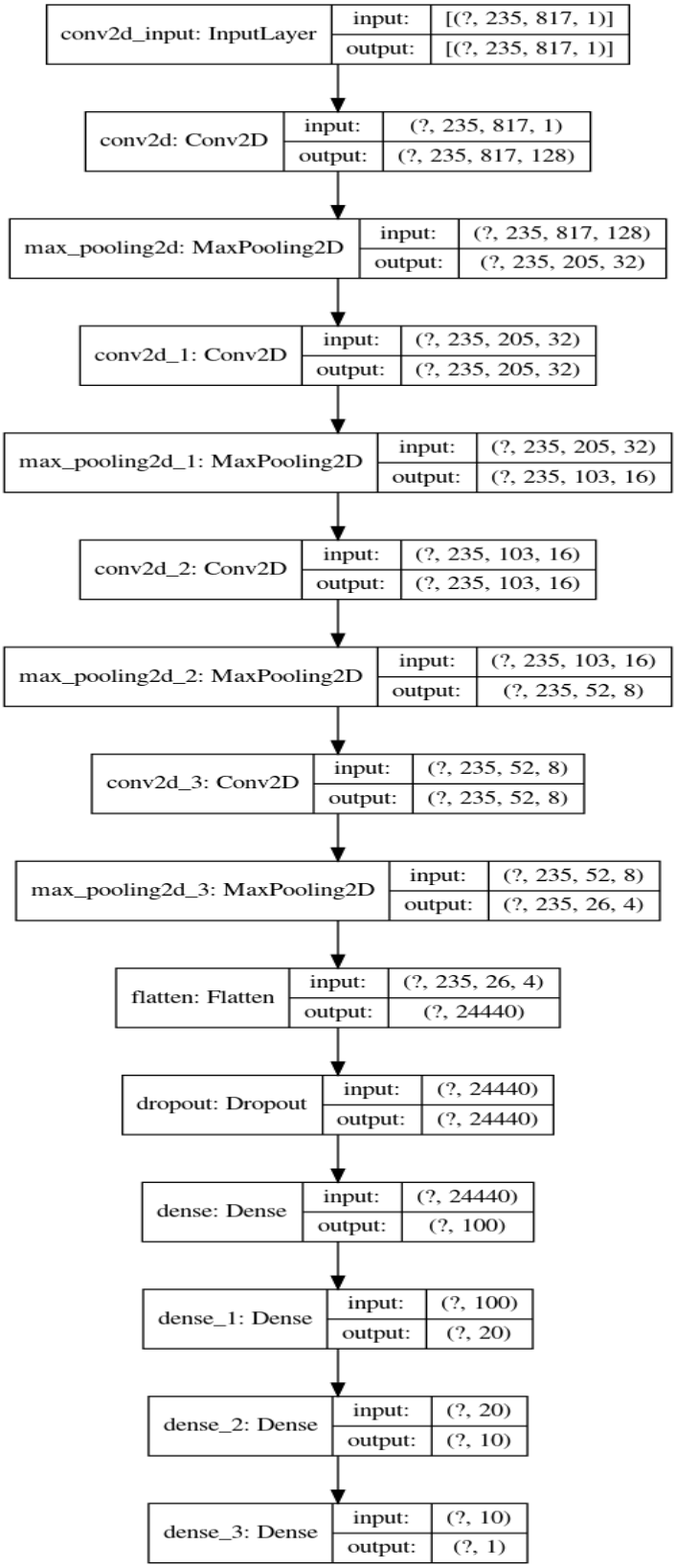

Figure 2. 2D CNN model

A hopping window, of a 10 minute length and a 30 second stride, was implemented on the physiological data. The data set was standardized along each feature vector (e.g., physiological signal type) by fitting a standard scaler such that the mean was zero, and the standard deviation was one. Each window was labeled with the participant Mackworth clock response accuracy. The multivariate physiological time series inputs were concatenated into a 2 Dimensional (2D) array, with the shorter time series being zero padded. The arrays were then reshaped such the number of rows and columns would be comparable. The resultant set of 2D arrays were randomly partitioned into a training and validation test, such that the validation set accounted for 20% of the data. A 2D Convolution Neural Network (CNN) regression model (Figure 2) was developed and trained on the input physiological data to predict the corresponding participant Mackworth clock response accuracy, on the training and validation set. In order to minimize the effect of outliers on the training, the loss function was chosen to be the Mean Absolute Error (MAE), rather than the Mean Squared Error (MSE).

The large difference in sample rates dictated the specific class of Deep Learning architectures selected. Initial attempts to use LSTM models were unsuccessful. Recurrent Neural Networks (RNN) such as LSTM models are frequently used for time series regression and classification tasks. However, an LSTM requires feature vector inputs of a uniform length, which given the varying sample rates, and resultant varying length time series could only be accomplished by either: 1) Zero padding the shorter time series, or 2) Having separate parallel LSTM models for each inputs of different length and then concatenating the LSTM outputs. Both of these approaches proved unsatisfactory.

In the second option, the number of LSTM hidden units has to be the same for each model to ensure that the outputs of the LSTMs are of the same length and can be concatenated. However, the selection of the number of LSTM hidden units depends on the length of the input feature vectors. Too few hidden units with respect to the input length will cause the under-fitting of the model. While too many hidden units will result in over-fitting. Due to large variation in input feature vector length an appropriate number of hidden LSTM units could not be determined and the model under-fit some inputs while over-fitting others, essentially ignoring certain inputs, or being too tightly coupled to others.

Zero padding the LSTM inputs was a viable solution; however, it was determined that CNN models heavily zero padded inputs models performed better than the LSTM. This is attributed to the fact that contribution of the zero padding in a CNN can be minimized by an appropriately chosen number of convolution and pooling layers, with appropriate kernel sizes.

Vigilance Task Difficulty/Cognitive Load Classification

In order to create a more difficult vigilance task the above described Mackworth clock was modified to have two, rather than one circulating dot (Figure 3). Both dots would move in a circle at regular increments and either dot would at random times experience a discontinuous jump. The participants have to monitor both dots and

acknowledge when either dot exhibited a discontinuous jump. The data preparation was identical to that described above, except that now each physiological signal window received a binary classification labeled corresponding to either, the one or two dot Mackworth clock test. The labels were on-hot encoded (i.e., [01] and [10]). A 2D CNN classifier was developed, tuned and trained to predict the correct binary classification label. The specific CNN model used was similar, but not identical to the one above (Figure 4). A binary cross-entropy loss function was used during training.

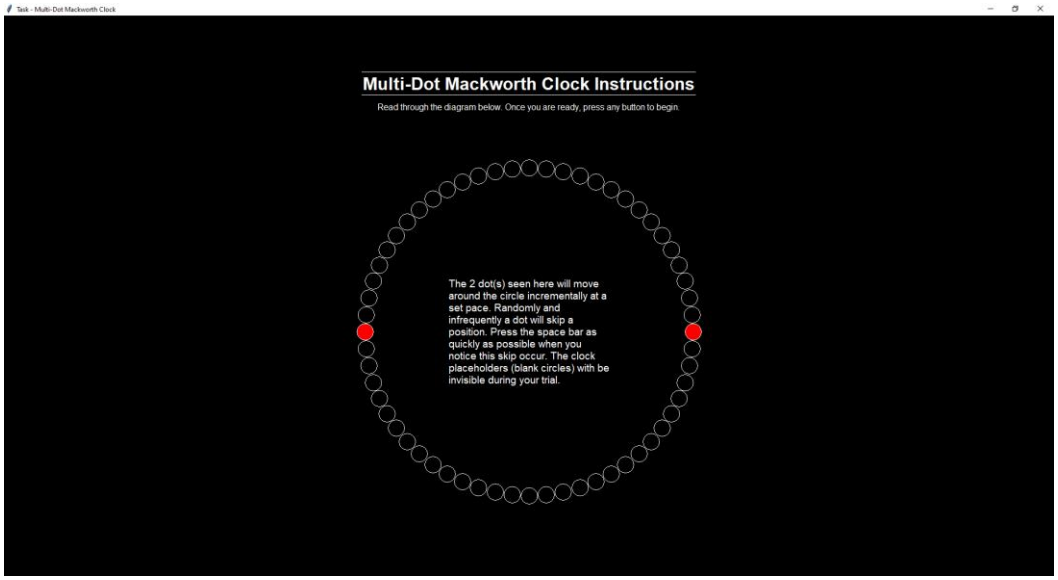

Figure 3. Overload Mackworth Clock Task Screen

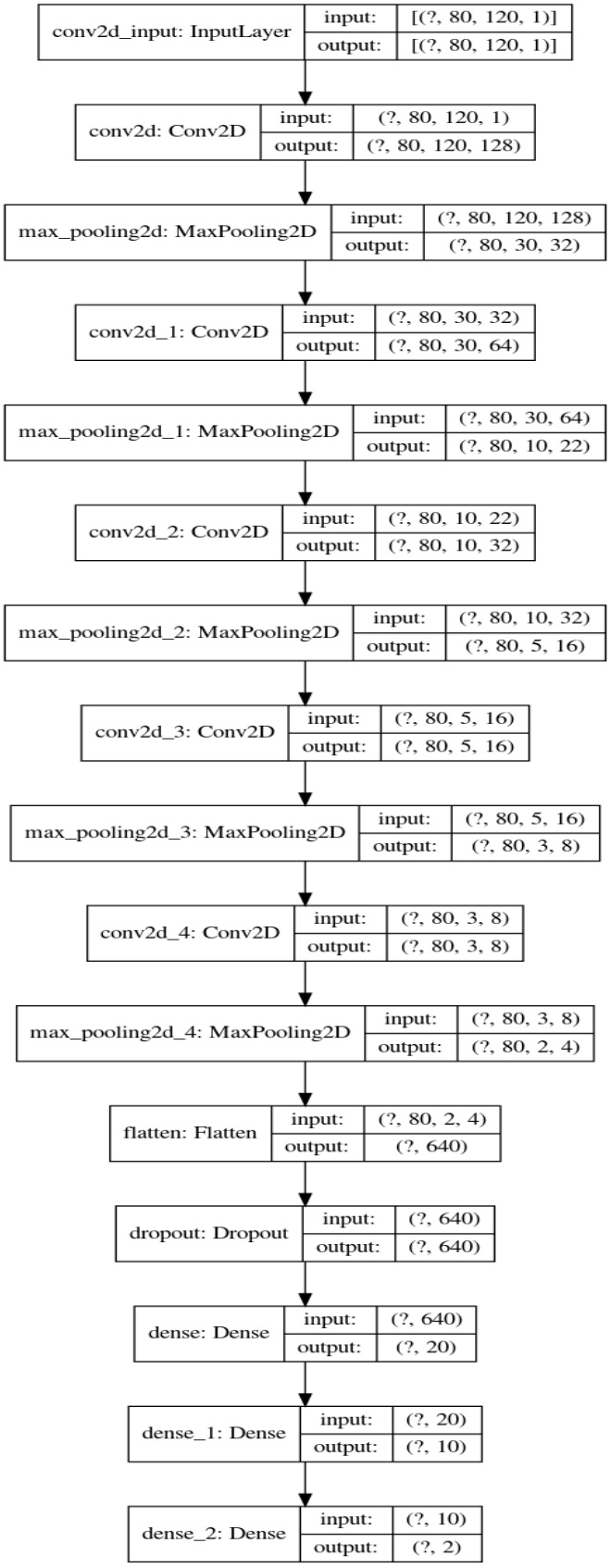

Figure 4. 2D CNN Model for Binary Classification

RESULTS/PROJECT SUCCESS

Nearly all participants showed performance decrement as expected which led to the assumption that the protocol successfully induced cognitive fatigue as intended. This is important because the first attempt in this study did not have conclusive performance decrements to label the dataset. Task performance decrement was relatively simple to identify and cognitive fatigue was inferred based on previous studies of the Mackworth clock. EEG was collected on all participants but was not used in algorithm development due the lack of confidence in the commercially available black box algorithm. Raw EEG signals collected will be filtered and added to our ML algorithm in a later phase of this effort. Subjective data was also collected but not used as an input to the algorithm.

The 2D CNN regression model described above was tuned and set to train for 200 epochs (e.g. iterations through the data). The plots of the training and testing histories, including the MAE and MSE are presented below (Figure 5). The plots suggest that the model was well behaved exhibiting no over, or under-fitting. The best-achieved model accuracy occurred on the 174st epoch. The model was able to predict the participant Mackworth clock monitoring accuracy on the validation set to within a MAE of 2.54%. For instance, if a user is responding with 80% accuracy in a given time window, the algorithm will output a predicted task performance accuracy for the given time window between 77.5% and 82.5%. This implies that cognitive performance on a simple task (the Mackworth clock) can be predicted using only physiological signals collected from a wrist-worn device. Most Warfighter tasking is not as simple as tracking a single dot on a projected path, but this provides sound evidence that further research needs to be pursued with more complex tasking.

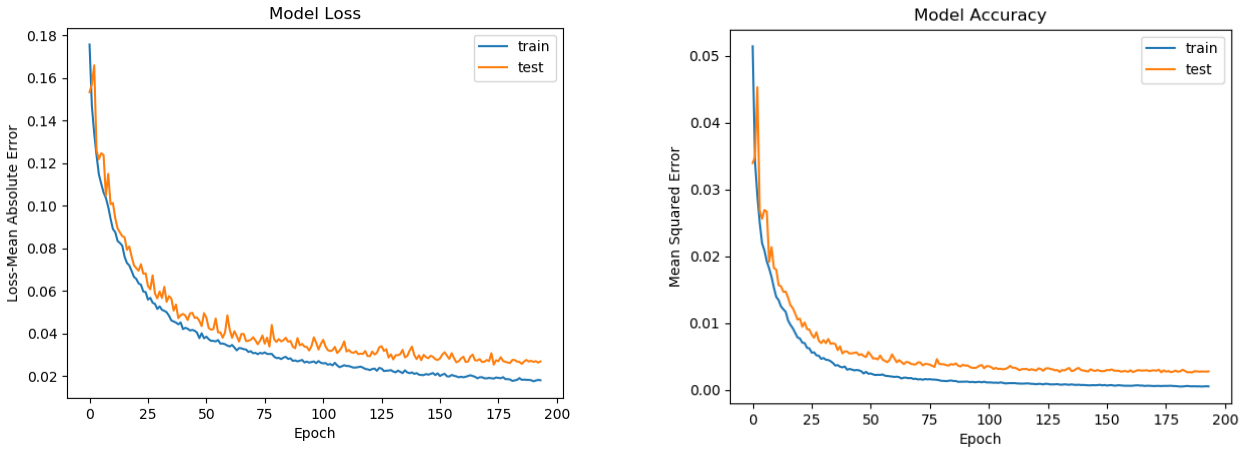

Figure 5. Training Plots for Mackworth Clock Performance Accuracy Model

The CNN binary task difficulty/cognitive load classifier model was likewise tuned and set to train on 200 epoch. The training and testing histories of both the binary cross entropy training loss function, as well as the binary classification accuracy is presented below (Figure 6). The model was well behaved, though did begin to show symptoms of overfitting in the last 20 epochs. The best achieved cognitive load/task difficulty binary classification accuracy on the validation set for was 86.57%, which translates to the algorithm being able to detect when the participant is performing the cognitively overloaded task using only the biometric signals collected from the wrist-

worn device. Although, this task is not a perfect representation operational relevant tasking, it provides useful insight that these physiological signals can be used in predicting and detecting cognitive performance. Further testing, with more complex tasking, is needed to prove the importance of using these physiological signals in cognitive state determination. Implications of successful cognitive state detection are limitless; especially, as human machine teaming research and development continues to advance. Real-time physiological signals could provide useful information for intelligent machines to adjust the modality and complexity of the tasking required by the operator for optimal performance in a given cognitive state.

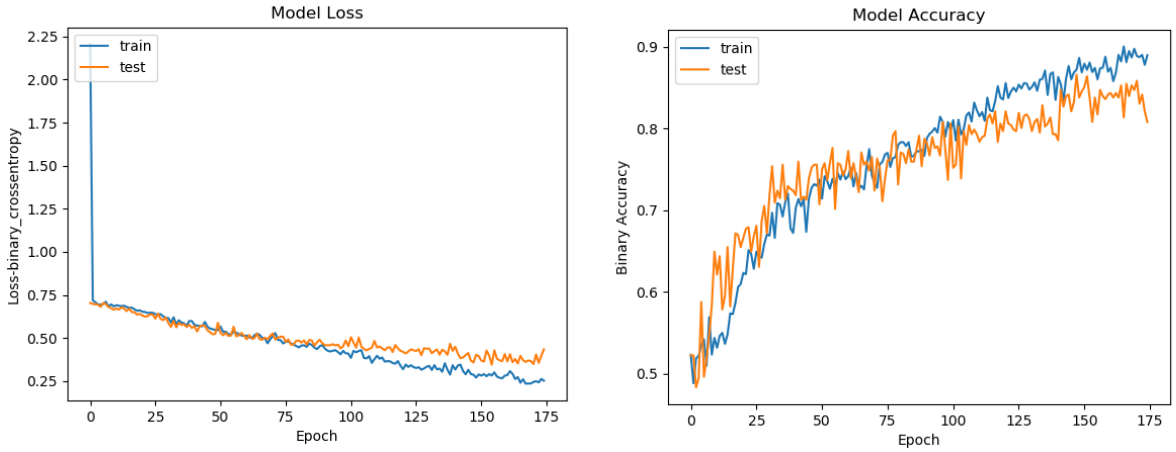

Figure 6. Training Plots for Binary Classification of Operator Workload

NEXT STEPS

This project has yet to transition, as it has been waiting for results to be published to show the full capability of the technology. That being said, a group funded by Marine Corps Warfighting Lab (MCWL) has expressed interest in the technology and we collected a large data set from a 48 hour field test event of an expeditionary medicine team. The event included a team of 9 Warfighters performing normal medical tasking (e.g., triage, surgery, wound care, etc.) for an operational window of 48 hours, as identified in the requirement. The Wearables team fit each Warfighter with wrist-worn devices to collect data during this event as the Warfighters will be faced with lack of sleep and other normal procedures that will induce fatigue. Periodically, the participants performed vigilance tasks for unambiguous labels of current cognitive state. The dataset will be used to develop a similar ML algorithm to detect performance decrements and other fatigue related occurrences. Additionally, the Commanding Officer (CO) and Executive Officer (XO) of the AEGIS Training and Readiness Center (ATRC), collocated on the Dahlgren campus, requested to be briefed on the results from their involvement (providing active duty participants for the lab testing) in the effort. This brief will be used to open doors to specific tasking that would be useful for Naval leadership to monitor operator performance, as active duty Warfighters provide valuable feedback to make a system that is needed and useful to the Navy.

TECHNICAL RISKS

The team planned for and mitigated all risks, but the SARS-CoV-2 (COVID-19) pandemic caused a significant delay in data collecting due the altered operating state of government facilities. Testing was paused, for approximately 3 months, while a new Standard Operating Procedure (SOP) was created and additional safety precautions implemented, to insure health and safety of all participants and researchers. The two major technical risks were developing a protocol that did not adequately induce fatigue and that the physiological data simply would not correlate with fatigue. To mitigate the protocol risk, the team performed a deep literature review, as well as, contacted a cognitive science from the Naval Health Research Center, Dr. Tim Dunn, who specializes in cognitive performance. Literature review and our collaboration with Dr. Dunn gave us confidence in the protocol developed. The second risk of there not being a correlation between fatigue response and physiological data is common in research. To increase the likelihood of success in finding a correlation in the dataset, the team researched physiological signals correlated with cognitive performance and chose a device that measured many of those signals. Both mitigations proved successful, as shown in the results above. There were various other smaller risks such as participant recruitment and having the IRB approve the protocol, but our team is experienced in these types of studies; therefore, navigating the IRB process and recruit were not an issue. The COVID-19 pandemic hit at the beginning of our testing phase and disallowed us to continue to test. Fortunately, the team quickly adapted the testing protocol in accordance with leadership direction to ensure the safety of the team and participants. Although there were delays due to the changing circumstances, the team still collected plenty of data to achieve acceptable results.

COLLABORATORS

The Dahlgren team included Brandon Marine, Rachel Sides, Igor Shtau, and Dr. Jessica Jones from V13, as well as multiple New Employee Developmental Assignment (NEDA) rotations (Rachel Fronzo, Jacob Gray, and Khade Grant). Additionally, test participants were recruited from all departments and were vital to this projects success. The main collaborator of this effort was with a cognitive scientist from Naval Health Research Center, Dr. Tim Dunn. Dr. Dunn provided valuable insight in inducing and measuring cognitive fatigue. This collaboration was simply information sharing; however, these promising results will lead to continued collaboration to develop proposals related to fatigue. Additionally, ATRC provided participants for the study. ATRC participants also provided valuable feedback regarding the use of wearable technology in Naval environments.

WORKFORCE DEVELOPMENT

This project supported four incoming NEDA rotations from various departments, and supported five team members to complete their external NEDA rotations. Additionally, knowledge learned during this project has helped two team members to pursue advanced degrees from the University of Virginia.

DELIVERABLES/BIBLIOMETRICS

1. Prototypes / Demonstrations / Awards
   - Prototype: 2D CNN model to predict Mackworth Clock performance accuracy
   - Prototype: 2D CNN model for binary classification of operator workload
   - Award: Kniffin, Alex. 2020 G. Dennis White Early Career Human Systems Integration Practitioner Award. Naval X, Alexandria, VA, July 2020.

2. Publications
   - Conference Presentation: Sides, R., Kniffin, A., Fatigue Detect/Prediction using ML and Wearable Technology. Virtual due to COVID-19, November 2020.
   - Government Report: Kniffin, A., Marine, B., Shtau, I., Sides, R., Jones, J., Fatigue Detection/Prediction Using Machine Learning and Wearable Technology. NSWCDD-TR-21-00058
   - White Paper: Kniffin, A., Strategic Physiological After Action Reporting Tool and Analysis (SPAARTA). August 2020.

3. Advanced Degrees
   - Advanced Degree In Progress: Kniffin, Alex. University of Virginia, Master of Science in Mechanical Engineering. Expected May 2022.
   - Advanced Degree In Progress: Sides, Rachel. University of Virginia, Master of Science in Chemical Engineering, Expected May 2024.

REFERENCES

1. L. Bi, R. Zhang and Z. Chen, "Study on Real-time Detection of Alertness Based on EEG," *2007 IEEE/ICME International Conference on Complex Medical Engineering*, Beijing, 2007, pp. 1490-1493, doi: 10.1109/ICCME.2007.4381994.
2. Bryant DCN, Deluca J. Objective Measurement of Cognitive Fatigue in Multiple Sclerosis. Rehabilitation Psychology. 2004;49(2):114–122. [Google Scholar]
3. Cho, Jungchan, and Hyoseok Hwang. "Spatio-Temporal Representation of an Electoencephalogram for Emotion Recognition Using a Three-Dimensional Convolutional Neural Network." *Sensors (Basel, Switzerland)* vol. 20,12 3491. 20 Jun. 2020, doi:10.3390/s20123491
4. Cimtay, Yucel, and Erhan Ekmekcioglu. "Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition." *Sensors (Basel, Switzerland)* vol. 20,7 2034. 4 Apr. 2020, doi:10.3390/s20072034
5. Dar, Muhammad Najam et al. "CNN and LSTM-Based Emotion Charting Using Physiological Signals." *Sensors (Basel, Switzerland)* vol. 20,16 4551. 14 Aug. 2020, doi:10.3390/s20164551
6. DeLuca J. Fatigue: Its Definition, its Study and its Future. In: DeLuca J, editor. Fatigue as a Window to the Brain. Cambridge (MA): MIT Press; 2005b. pp. 319–325. [Google Scholar]
7. Henelius, Andreas et al. "Heart rate variability for evaluating vigilant attention in partial chronic sleep restriction." *Sleep* vol. 37,7 1257-67. 1 Jul. 2014, doi:10.5665/sleep.3850
8. Holtzer R, Foley F. The relationship between subjective reports of fatigue and executive control in multiple sclerosis. Journal of Neurological Sciences. 2009;281(1–2):46–50. [PMC free article] [PubMed] [Google Scholar]
9. Holtzer R., Shuman M., Mahoney J.R., Lipton R., Verghese J. Cognitive Fatigue Defined in the Context of Attention Networks. Aging Neuropsychol. Cogn. 2010;18:108–128. doi: 10.1080/13825585.2010.517826. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
10. Jung TP, Makeig S, Stensmo M, Sejnowski TJ. Estimating alertness from the EEG power spectrum. IEEE Trans Biomed Eng. 1997 Jan;44(1):60-9. doi: 10.1109/10.553713. PMID: 9214784.
11. 23Lichstein, K. L., Riedel, B. W., & Richman, S. L. (2000). The Mackworth clock test: A computerized version (statistical data included). The Journal of Psychology.
12. Liu, Ning-Han et al. "Recognizing the degree of human attention using EEG signals from mobile sensors." *Sensors (Basel, Switzerland)* vol. 13,8 10273-86. 9 Aug. 2013, doi:10.3390/s130810273
13. 22Mackworth NH (1948) The breakdown of vigilance during prolonged visual search. Quarterly Journal of Experimental Psychology 1: 6–21 10.1080/17470214808416738 [CrossRef] [Google Scholar]
14. Makeig S, Inlow M. Lapses in alertness: coherence of fluctuations in performance and EEG spectrum. Electroencephalogr Clin Neurophysiol. 1993 Jan;86(1):23-35. doi: 10.1016/0013-4694(93)90064-3. PMID: 7678388.
15. Makeig S, Jung TP. Changes in alertness are a principal component of variance in the EEG spectrum. Neuroreport. 1995 Dec 29;7(1):213-6. PMID: 8742454.
16. Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ. Dynamic brain sources of visual evoked responses. Science. 2002 Jan 25;295(5555):690-4. doi: 10.1126/science.1066168. Erratum in: Science 2002 Feb 22;295(5559):1466. PMID: 11809976.
17. Matousek M, Petersen I. A method for assessing alertness fluctuations from EEG spectra. Electroencephalogr Clin Neurophysiol. 1983;55:108–113. [PubMed] [Google Scholar]
18. 24McAvinue, L.P., Habekost, T., Johnson, K.A. *et al.* Sustained attention, attentional selectivity, and attentional capacity across the lifespan. *Atten Percept Psychophys* 74, 1570–1582 (2012). https://doi.org/10.3758/s13414-012-0352-6
19. Pattyn N, Neyt X, Henderickx D, Soetens E. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? Physiol Behav. 2008 Jan 28;93(1-2):369-78. doi: 10.1016/j.physbeh.2007.09.016. Epub 2007 Oct 3. PMID: 17999934.
20. Schwid SR, Tyler CM, Scheid EA, Weinstein A, Goodman AD, McDermott MP. Cognitive fatigue during a test requiring sustained attention: a pilot study. Multiple Sclerosis. 2003;9(5):503–508. [PubMed] [Google Scholar]
21. Siddle DA. Vigilance decrement and speed of habituation of the gsr component of the orienting response. Br J Psychol. 1972;63(2):191–194. [PubMed] [Google Scholar]
22. Subasi, Abdulhamit. (2005). Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients. Expert Systems with Applications. 28. 701-711. 10.1016/j.eswa.2004.12.027.

23. Vuckovic, Aleksandra & Radivojevic, Vlada & Chen, Andrew & Popović, Dejan. (2002). Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. Medical engineering & physics. 24. 349-60. 10.1016/S1350-4533(02)00030-9.
24. *Wang X, Piñol RA, Byrne P, Mendelowitz D. Optogenetic Stimulation of Locus Ceruleus Neurons Augments Inhibitory Transmission to Parasympathetic Cardiac Vagal Neurons via Activation of Brainstem α1 and β1 Receptors. . 2014 Apr 30;34(18):6182–9. [*Google Scholar]