Comparing Deep Neural Network Architectures using Transfer Learning to Predict Cognitive States

Nicholas J. Napoli, Mudit Paliwal, Victoria R. Rodrigues
Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611, USA
Human Informatics and Predictive Performance Optimization (HIPPO) Lab, Gainesville, FL, 32611, USA

Angela Harrivel, Kellie D. Kennedy, Chad L. Stephens
NASA Langley Research Center, Hampton, VA 23681

I. Introduction

One of the ways we can measure the pilot operator's functional state is by monitoring and fusing physiological sensing data. Monitoring the functional state of these operators will help in improving autonomous feedback systems to improve aviation safety. These state of the art adaptive systems that aid the operator could recognize the cognitive states and can alert operators of catastrophic situations to aid them to prioritize their actions in order to prevent and mitigate emergency situations [1]. Improving our understanding of cognitive states using hybrid predictive models will help us decipher the appropriate adaptive aiding system to implement. However, physiological sensing data produces a lot of uncertainty in the predictive model's performance which is caused by noise from various sources, is susceptible to inter-subject differences, and is hard to characterize. To explore this uncertainty in the predictability of cognitive states using noisy data, it becomes essential to differentiate the untapped potential of different learning models. This can help us navigate a solid foundation on our research towards the increase of predictability of these states and mitigate undesired multi-modal parameters.

Prior Work: The operator's mental state is not directly inferred but is learned through human physiological and behavioral data. There are different modalities that are used to capture this data such as eye movement, electrocardiogram (EEG), electroencephalogram (ECG), heart rate, galvanic skin conductance, and respiration rate. These modalities can be used individually or in conjunction (Fused) to predict cognitive states. However the identification of subject variability is difficult for one machine learning model's decision boundary to separate [2, 3]. Understanding the physiological sensing data and their differences running through different modalities can increase the discernment of different cognitive states explicitly and eventually improve training for pilots and vehicle operators [4, 5].

In recent years, deep learning has attained great popularity and its ability to extract features based on visual learning (e.g., convolutional neural networks) has accelerated its growth in different domains [6–9].These methods have dramatically improved state-of-the-art speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics [6]. In this context, deep learning could significantly simplify the process pipelines by allowing end-to-end automation of pre-processing, feature engineering, and classification modules, while also reaching competitive performance on the target task. There are different building blocks of the neural network architecture and information of different modalities can be formatted and fed into a deep learning model. While there is much research on using deep learning architectures to classify physiological sensing data, there is little work that focuses on comparing different model architectures and how deep learning performs in parallel to machine learning models on the same set of data for cognitive state detection [2, 10]. Also, the vulnerability of models based on low numbers of samples (i.e subject's data) is not clear, as is how it plays an important role in accuracy and stability. Improvements would offer better generalizations based on new random data from uncontrolled environments [11].

Challenges: Research to date has focused on predicting cognitive states on a very limited set of data and a low amount of cognitive classes to detect (e.g., only high and low workload) [2, 12, 13]. The availability of large data sets containing vast amounts of sample data for training is often mentioned as one of the core components of any machine learning model. It is therefore hard to generalize models with physiological sensing data, given the relatively high cost of data collection [10]. Additionally, the high dimensionality of this data makes it difficult to estimate the decision boundaries, which requires an increased amount of data to reduce the confidence boundary of the estimated parameters, regardless of the type of ML approach (e.g., deep learning, classical approaches) [2, 10].

Insights: To overcome the shortcomings of having insufficient data, two similar psychophysiological experimental design sets have been combined. These were developed at NASA Langley's Research Center, and have been referred to as Scenarios for Human Attention Restoration using Psychophysiology (SHARP) [14]. In order to explore the weakness of insufficient data, we utilize data augmentation (e.g., a method for up-sampling) techniques to create new data examples by artificially generating them from the existing data. Adding more training examples would help us to utilize more complex parameters while generalizing the model and reducing over fitting [11].

Data augmentation has been recently used and much of the research has not been fully utilized to look for interestingly different modal parameters that are influenced by the greater number of training samples [10]. About 40% of the research up to date using deep learning architectures has used convolutional neural networks (CNN) and 13.5% have utilized Recurrent Neural Networks (RNN) [10]. While not limited to deep learning models, there has been little work which brings out the fusion of the two neural network architectures models. Lastly, the need to extract features is completely removed in deep learning models. By leveraging the deep learning design, we can indeed use raw physiological sensing data more often as input and avoid the task of extracting features from the data.

Contributions: Through our insight of utilizing multi-modal physiological sensing and machine-learning/deep-learning architectures, we will put forward the following premise:

1) We will show how these Cognitive Models improve upon the use of classical machine learning approaches when combining data sets (SHARP-1 and SHARP-2).
2) We will show how Deep Neural Network architectures (e.g, transfer learning, data augmentation) can improve performance compared to other ML methodologies.

II. Methods

A. Experimental Design

The *SHARP-1* dataset contains twenty four commercial aviation pilots who participated in the study to carry out the tasks in a motion-based flight simulator. The tasks were designed to induce attention-related human performance-limiting states (AHPLS) and low/high workload conditions(LW, HW) in experimental flight scenarios. While performing the task the pilots were asked to wear psychophysiological sensors to collect physiological sensing data. The study was approved by the Institutional Review Board of the NASA LaRC and under the consent of the participants. There were different scenarios designed to create attentional demand, startle/surprise (SS) and mental workload throughout the experimental scenarios to induce AHPLS. Apart from this experiment, there were subjective self-report questionnaires (NASA TaskLoad Index, AHPLS experienced questionnaire, and NASA Situation Awareness Rating Technique) after each task to assess situation awareness. The cognitive states in AHPLS include Baseline Rest, Tetris for Channelized attention, MATB (Multi-Attribute Task Battery: available at http://matb.larc.nasa.gov/) for High and Low Workload and a movie scene observation for startle/surprise which is similar to the experiments done in SHARP-2. These benchmarking tasks have has been discussed previously [15].

The *SHARP-2* dataset has a similar experimental design as SHARP-1. This data set contains data from ten commercial aviation pilots who were asked to perform the same AHPLS task to induce the experimental scenario as defined above in the SHARP-2 design. The data collection and study design of the two methods slightly differ in the sense of the instruments used to collect multiple modalities, which will be discussed in the next section.

B. Data Collection and Study Design

The *SHARP-1* dataset has a sampling rate for the recording of electroencephalogram data which was 256 Hz using Advanced Brain Monitoring's B-Alert X24 electrode head net. The Mind Media, B.V. Nexus Mark II system was used to measure three-point electrocardiogram (ECG), respiration, and galvanic skin response (GSR) signals. The three above-mentioned modalities were also recorded using a sampling rate of 256 Hz. Using a Smart Eye eye-tracking system, the pilot's visual behavior and eye responses were recorded in the flight simulator.

The *SHARP-2* dataset utilizes the B-Alert X24 to collect electroencephalogram data at a frequency of 256 Hz. Additionally for SHARP-2, electroencephalogram data was recorded using the Muse by InteraXon at 500 Hz. The Empatica was used to collect blood volume pulse at 24 Hz, galvanic skin response and skin temperature at 4 Hz, heart rate at 1.3 Hz, and acceleration of the sensor at 32 Hz. The Muse by InteraXon was used to collect electroencephalogram data at 500 Hz and acceleration of the sensor at 50 Hz. The NeXus-10 Stream was used to collect electrocardiogram and respiration data at 256 Hz. The Spire collected respiration force and acceleration of the sensor at 473 Hz. The Tobii 2 Eye Tracker collected gaze location as well as left and right pupil diameter at 50 Hz.

All measured time series were recorded using MAPPS (EyesDx, Inc., Coralville, IA), a software suite designed to collect aircraft and simulator state, event markers, video, and pilot physiological and behavioral data. The software time-synchronizes all data channels for real-time review, as well as post-hoc analysis.

C. Algorithmic Design and Data Analysis

RQ1: How do these Cognitive Models improve using classical machine learning approaches when combining data sets (SHARP-1 and SHARP-2).

Considering the physiological sensing data acquired from two different experimental designs, it is very important to consider first the equipment it was captured from. Secondly, we look at the bench mark task that were used to induced the different cognitive states (CA, HW, LW, SS). Thirdly, it is important that task were carried out with instructions under the same environment involving the same task time and breaks. All these conditions will lead to a successful conjunction of subjects from two data sets. The next challenge is to bring the shape and dimensions of the data sets together. This can be done by re-sampling the sensor data to the provided frequency while interpolating when necessary using a linear method and then creating a single frame for one modal sensor. Different data inputs will be structured into raw EEG signals with minimum pre-processing, and EEG signals with some feature extraction before feeding into the model. Henceforth, the machine learning models used in the previous design [16, 17] will be used to compare the performance of classification of multiple sates using multiple sensing modalities in flight simulators.

RQ2: Are there Deep Neural Network Architectures (e.g, transfer learning, data augmentation) which Improve Performance Compared to other ML methodologies?

We hypothesize that the approach of using Deep Learning methodologies gives us three potential benefits for cognitive state detection. Firstly, deep learning architecture (e.g., transfer learning) can provide an initial model structure that is pre-developed which can be refined (i.e., weights of the NN) to our existing dataset. Secondly, the computational time to train to reach an optimal decision is reduced for faster training times. Lastly, data augmentation methods can also improve the number of samples for the model to train to continue to improve the confidence internal of the decision boundaries. Using the concept of transfer learning, we will explore different architectures such as CNN and RNN to uncover how these methods increase classification accuracy for predicting cognitive states. With this experimental setup, different physiological sensing data sets will be fed into the architecture for classification of the four cognitive states. Therefore different combinations of inputs will help us examine and compare the need to extract features when working with deep learning architectures. The benefits of utilizing Neural Networks architecture such as CNN is that the feature engineering is performed by the model, unbiased by the analyst's ability to develop features. Consequently, it will give us insight in understanding on how neural network models perform when compared to classical classification methods that require feature engineering for cognitive state detection. This addresses whether there is truly hidden or embedded information within the signal that the research community is not aware of that provides predictive lift to the algorithm. Apart from different data representations, the hyper-parameters of various deep learning architectures will be examined for the development of the best learning model. Above mentioned models will then be evaluated based on predictive accuracy, precision, and recall. These models will include leave-one-subject-out cross validation in order to be pragmatic towards our goal of achieving inter-subject classification. These models will then be compared to the classical machine learning methodologies used in RQ1 putting forward an extensive quantitative comparison of the different model's performance.

We believe that deep learning methodology will provide a clear insight into the pattern recognition of the physiological sensing data. This will bring forward the utilization of feature engineering and how different architectures perform in a dynamic environment, which is required when developing models for workable adaptive aiding system.

III. Conclusion and Future Work

This study seeks to explore the methodologies of cognitive state detection associated with human physiological responses and visual behaviour. This understanding can reveal the hidden dynamics of pattern recognition in the physiological sensing data to potentially advance towards improving adaptive aiding system. In addition, this research will leverage a solid foundation for designing autonomous human machine integration for real-time classification of human state predictions. This work will quantify the comparison of two fields: first predicting multi-class cognitive state prediction using classical machine learning algorithms, and second understanding the need of different deep learning models to improve performance and discover the unbiased need of feature engineering. Future work will seek to incorporate fusion of models together to generate the results in an uncontrolled environment. In order to implement this future work, we need to create experimental designs which are able to induce all the cognitive states in a single task replicating the actual events of the flight scenario. As more efficient models are developed in the controlled test environment, they can deployed and evaluated in adaptive aiding systems in the uncontrolled environment.

References

[1] Wilson, G. F., and Russell, C. A., "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human factors*, Vol. 45, No. 4, 2003, pp. 635–644.

[2] Fahimi, F., Zhang, Z., Goh, W. B., Lee, T.-S., Ang, K. K., and Guan, C., "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *Journal of neural engineering*, Vol. 16, No. 2, 2019, p. 026007.

[3] Wei, C.-S., Lin, Y.-P., Wang, Y.-T., Lin, C.-T., and Jung, T.-P., "A subject-transfer framework for obviating inter-and intra-subject variability in EEG-based drowsiness detection," *NeuroImage*, Vol. 174, 2018, pp. 407–419.

[4] Team, C. A. S., "SE211: Airplane State Awareness - Training for Attention Management (R-D)," , Mar. 2015. URL `http://www.skybrary.aero/index.php/SE211:_Airplane_State_Awareness_-_Training_for_Attention_Management_(R-D)`.

[5] Team, C. A. S., "Airplane State Awareness Joint Safety Analysis Team Interim Report," , Mar. 2015. URL `http://www.skybrary.aero/index.php/Commercial_Aviation_Safety_Team_%28CAST%29_Reports`.

[6] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *nature*, Vol. 521, No. 7553, 2015, pp. 436–444.

[7] Graves, A., Mohamed, A.-r., and Hinton, G., "Speech recognition with deep recurrent neural networks," *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 6645–6649.

[8] Zeiler, M. D., and Fergus, R., "Visualizing and understanding convolutional networks," *European conference on computer vision*, Springer, 2014, pp. 818–833.

[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[10] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J., "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, Vol. 16, No. 5, 2019, p. 051001.

[11] Wang, F., Zhong, S.-h., Peng, J., Jiang, J., and Liu, Y., "Data augmentation for eeg-based emotion recognition with deep convolutional neural networks," *International Conference on Multimedia Modeling*, Springer, 2018, pp. 82–93.

[12] Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., and Zuo, S., "EEG-based spatio–temporal convolutional neural network for driver fatigue evaluation," *IEEE transactions on neural networks and learning systems*, Vol. 30, No. 9, 2019, pp. 2755–2763.

[13] Appriou, A., Cichocki, A., and Lotte, F., "Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals," *IEEE Systems, Man and Cybernetics Magazine*, 2020.

[14] Stephens, C., Harrivel, A., Prinzel, L., Comstock, R., Abraham, N., Pope, A., Wilkerson, J., and Kiggins, D., "Crew state monitoring and line-oriented flight training for attention management," 2017.

[15] Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., and Pope, A. T., "Psychophysiological sensing and state classification for attention management in commercial aviation," *AIAA Infotech@ Aerospace*, 2016, p. 1490.

[16] Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., and Pope, A. T., "Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing," *AIAA Information Systems-AIAA Infotech@ Aerospace*, 2017, p. 1135.

[17] Napoli, N., Adams, S., Harrivel, A. R., Stephens, C., Kennedy, K., Paliwal, M., and Scherer, W., *Exploring Cognitive States: Temporal Methods for Detecting and Characterizing Physiological Fingerprints*, ???? doi:10.2514/6.2020-1193, URL `https://arc.aiaa.org/doi/abs/10.2514/6.2020-1193`.