Using Machine Learning to Predict Reading Strategies from fNIRS Data

Matthew Campbell[1], Indika Kahanda*[2]

[1] School of Computing, University of North Florida, Jacksonville, United States.
n00974082@unf.edu

[2] School of Computing, University of North Florida, Jacksonville, United States.
Indika Kahanda, 0000-0002-4536-6917

*corresponding author

Keywords: fNIRS, Machine Learning, Reading Strategies.

Abstract. Functional Near-Infrared Spectroscopy (fNIRS) is an emerging Neuroimaging technology that is useful to researchers due to its low cost and flexibility. In this study, we develop machine learning models for predicting the strategies used by readers. Specifically, we develop supervised learning models based on Support Vector Machines and Artificial Neural Networks for predicting 18 different reading strategies using fNIRS data gathered from 103 participants who are asked to read a text passage. Our experimental results based on F1 values indicate that this approach is overall promising and that fNIRS data could potentially be used to predicting some of the strategies used when reading. For example, the "using text headings" strategy displayed a relatively high performance of 69%. This is the first study that develops machine learning models for fNIRS data from a reading exercise. Therefore, our findings have implications for both the machine learning application researchers and educators.

1 Scientific Background

Functional near-infrared spectroscopy (fNIRS), which is a recently developed non-invasive technique for functional monitoring of the brain, works by shining near-infrared light using diodes facing the skull [1]. Once the light goes through the skull, the light bounces back or gets absorbed depending on the ratio of the oxygenated hemoglobin and the deoxygenated hemoglobin [1]. One crucial aspect about hemoglobin is that the deoxygenized hemoglobin and oxygenized hemoglobin have different optical spectrums, allowing the oxygen levels to be recorded [1]. Typically, fNIRS experiments have an applicability rate of approximately 78% whenever a particular part of the brain is examined [2].

Since fNIRS technology is more recent, applications of machine learning techniques for fNIRS data are only a handful in which the majority focused on predicting the mental workload [3, 4, 5, 6, 7, 8, 9]. Benerradi and others [3] explored the feasibility of predicting mental workload using fNIRS data. Specifically, they evaluated three learning models' performance: logistic regression, Support Vector Machines (SVMs), and Convolutional Networks (CNNs), on two tasks that involve clicking moving circles on a computer screen. They used features based on the Correlation Signal Based Improvement (CBSI) mean, which is the average of four values in a channel. They report accuracies in the range of 46% to 75% [3].

Zhang and others [4] collected data about the gaze of drivers and various aspects of the driving task. For instance, the system saved the data on the direction of the drivers' gaze, pupil diameter, and lane position. Afterward, they used a decision tree algorithm to classify if the mental workload was high or low-stress. Audrey Girouard and others [5] examined the passive brain-computer interface, which is an interface that does not require any human interaction to record the data. This causes the brain to be another resource of data collection. The study's primary purpose was to check if a real-time classification of the data would be viable. As it turns out, even though the performance of the real-time analysis was lower than the performance of the prerecorded analysis, the accuracy of the real-time analysis was 85%, demonstrating that fNIRS analysis can be performed in real-time.

Sassaroli and others [6] have utilized fNIRS on the front of the brain to differentiate the amount of mental workload into three different categories. To legitimize fNIRS as a method for classifying different mental workload levels, the team had used a 3-nearest neighbor algorithm using the data that came from the levels of the oxygenated and deoxygenated hemoglobin. The oxygen levels relate to the memory tasks that the participants had to perform. The accuracies of the classification had a success rate anywhere from 44% to 72%, which is relatively high compared to random data, which had a success rate of 19%.

Daniel Afergan and others [7] experimented with having participants play a game where they had to guide computer-simulated autonomous air crafts using a path that consisted of straight lines. The participants also had to consider the obstacles on the screen, which incurred a hefty penalty if an aircraft wandered into them. The authors used additional air crafts or removed any excess air crafts to change the task's difficulty based on the participants' mental state. As a result, the researchers could reduce the error rate by 35% compared to the initial analysis. This showed that fNIRS could be used to augment the task performance of a participant utilizing dynamic difficulty.

Girouard and others [8] attempted to find ways to measure the data that comes from fNIRS measurements that tested the user's spatial memory. The spatial memory task was split into a series of subtasks with two different categories, one with high demand and the other with low demand. A classifier successfully differentiated the categories of the mental workload (about a 70% to 80% success rate). This and other early research projects helped illuminate how machine learning can be used with fNIRS data.

Yuskel and others [9] conducted an experiment where several pianists were chosen to test if fNIRS was used to provide a system of dynamic difficulty. Once the person has a relatively low cognitive workload, the system goes on to a more challenging version of the activity since they could perform the current version quite sufficiently. The researchers found that the participants can play more correct notes with the proper timing and play at a higher tempo than the control group. The participants felt like the transition to the greater difficulty took place at the correct time. This demonstrated that fNIRS could be used to break down more difficult jobs into easier jobs.

In our current study, we used machine learning techniques for fNIRS data generated from a reading exercise. We intended to explore the feasibility of predicting different reading strategies using supervised machine learning models based on the fNIRS features. The motivation comes from the fact that identifying different strategies used by a reader on the fly could potentially become a valuable tool for educators in the future. This is the first study that deals with applying machine learning techniques for fNIRS data related to a reading exercise to the best of our knowledge.

2 Materials and Methods

2.1 Data

For our experiment, participants were asked to read a passage about aliens. The readers were placed into three categories, which are *silent reading*, *reading aloud*, and *thinking aloud*. The thinking aloud group had to mention the strategies they were using verbally during the reading process. The other two groups were asked to indicate the strategies they used post-reading. All groups had to indicate to what extent they used a

Table 1: Summary statistics for the three reading groups. Units are in bars.

Measure/Group	Silent Reading	Reading Aloud	Thinking Aloud
Mean	$3.68e$-6	$4.81e$-6	$6.89e$-6
Median	$4.58e$-7	$9.43e$-8	$-2.93e$-6
Standard Deviation	$3.89e$-4	$4.41e$-4	$5.46e$-4
Minimum Value	$-8.67e$-3	$-7.78e$-3	$1.31e$-2
Maximum Value	$8.77e$-3	$7.80e$-3	0.121
Number of Participants	27	25	23
Avg Data Points Per Person	59,671	73,804	77,658

strategy with a 0-100 rating score. The fNIRS equipment monitored thirty-eight locations in ten regions of their brains for measuring the brain oxygen levels. The data was taken from 75 participants for this analysis, which generated approximately 5.7 million data points. The silent reading group had 1.6 million observations, while the reading aloud group and the thinking aloud group had 1.9 million and 2.1 million observations. Summary statistics are listed in the table below (see Table 1).

Table 2 lists the 18 strategies and the number of responses that were used for the downstream analysis. The missing data was approximately 5% (hence the varying number of responses for each category). For each participant, we added a binary class label representing each strategy. These labels were created using the means of the rating values recorded for each strategy across all responses (i.e., values below the mean were considered negative). The Table 2 also shows the number and the percentage of positive labels for each class label (i.e., strategy). Based on the overall mean values across participants (data not shown), the strategy that had the highest mean is “questioning what is in the text” (no. 6). On the other hand, “skimming” (no. 3) and “predicting” (no. 5) were the strategies least used.

2.2 Models and Setup

As mentioned before, we used two machine learning classifiers: Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs). Specifically, we developed 18 separate binary classifiers for each reading strategy. We used the average of the 38 channel values as the features (i.e., 38 features). SVMs utilized a C value of 1.0 and an epsilon value of 0.1. It also used a sigmoid function as its kernel. The algorithm also used a numerical tolerance of 0.001 and an iteration limit of 100. As for Neural Networks, it consists of a single hidden layer of 100 units. The activation method used was the Rectified Linear Unit. It makes use of the Adam optimizer. Alpha value was 0.0001. The maximum number of iterations was set to 200. It uses replicable training. Both models were implemented using the Orange Data Mining Tool[1].

The data points with missing feature values (the means of the brain activity data in a particular region) were removed from the analysis. The questions that did not have a response for a specific strategy were also removed. For instance, if a participant answered with a 0 on “Rereading” and did not answer on “skimming”, that participant’s response would be used in rereading, and the response in skimming would be discarded. As a result, the number of data points that correspond to the strategies was unequal (see Table 2). We use stratified 20-fold cross-validation for the evaluation. F1 was used as the performance metric.

Table 2: Statistics on the responses for each strategy.

No.	Strategy	# Responses	# Pos.	% Pos.
1	Rereading	72	42	58.3
2	Adjusting your reading rate	72	51	70.8
3	Skimming	58	36	62.1
4	Guessing the meaning of a word	71	42	59.2
5	Predicting	56	31	55.4
6	Questioning what is in the text	68	47	69.1
7	Arguing with the text	51	26	51.0
8	Rehearsing or repeating part of the text	68	51	75.0
9	Using text features such as headings or titles	51	25	49.0
10	Restating to yourself what is in the passage	62	43	69.4
11	Making connections between the passage	70	42	60.0
12	Made interpretations about the text	65	46	70.8
13	Elaborated to yourself about the text	52	35	67.3
14	Evaluated how well you understood the text	73	48	65.8
15	Evaluated how interesting the text was to you	67	47	70.1
16	Evaluated the quality of the text	60	38	63.3
17	Evaluated how difficult the text was for you to read	66	44	66.7
18	Evaluated whether you agreed with the text or not	59	43	72.9

3 Results

Overall, one of the primary observations that can be made is that both models have mixed performances across strategies (see Figure 1). Neural Networks performed slightly better in the F1 metric (0.44 and 0.41). This seems contradictory because neural networks usually require a lot of data to perform well, and SVMs perform well with smaller data sets. As for the individual strategies, strategy no. 18 (“evaluated whether you agreed with the text or not”) has the best performance from SVMs while no. 16 (“evaluated the quality of the text”) has the best F1 for ANNs. On the other hand, strategy no. 5 (“predicting”) showed the lowest F1 values concerning both SVMs and ANNs.

4 Conclusion

fNIRS is a relatively new technology that is gaining popularity due to its ease of use. The primary motivation of this study was to examine fNIRS-based machine learning models that can be used to distinguish whether a particular strategy is being used or not during reading. SVMs and ANNs were used to classify a set of binarized data points, which came from a reading exercise where each participant provided a score between 0 and 100 on how much they used a particular strategy. While both models performed just about equally overall, there are several strategies for which both models provided relatively high F1 values suggesting that machine learning models may be able to predict specific strategies effectively. Our F1 values ranged from 15%-69%.

As for the possible future work, there are several promising directions. Firstly, it would be interesting to explore incorporating the temporal aspects of the data. For example, instead of using the mean values of each channel, features could be developed by feeding the entire sequence of values. This would be suitable for a two-dimensional convolutional neural network. Alternatively, this could be provided into a recurrent neural network by concatenating the sequences of values for each channel. Secondly, we’d like to model this as a regression problem instead of binarized labels. While an ablation study involving individual reading groups would be enlightening, the number of partic-

[1]https://orangedatamining.com/

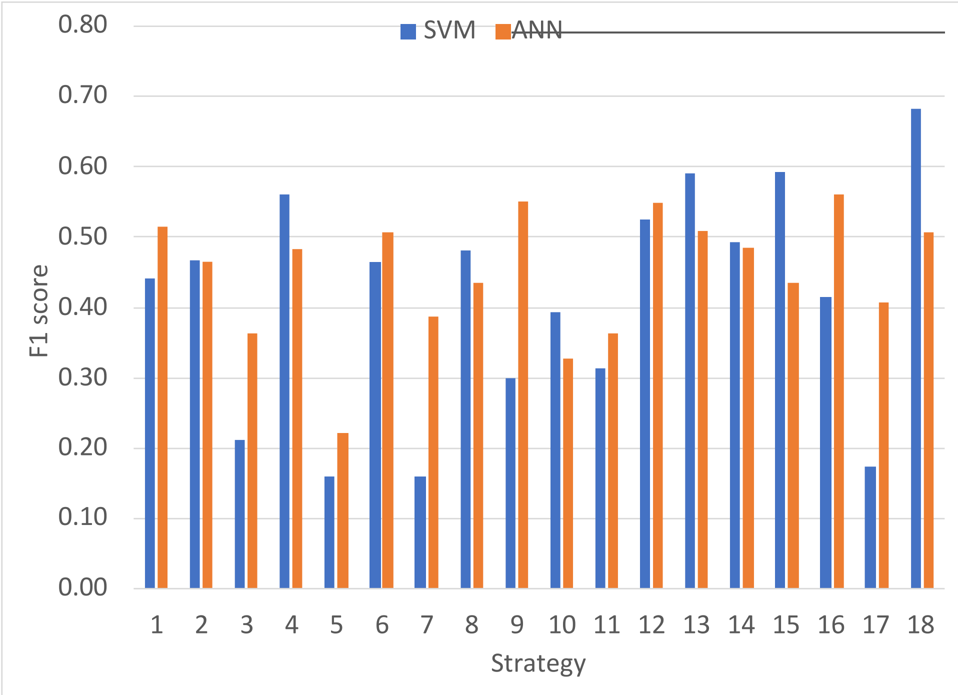

Figure 1: Performance comparison for SVMs versus ANNs for the 18 reading strategies using F1 value.

ipants for each group is seemingly too few in our current experiment. Hence, collecting more data with a larger group of readers is on the horizon. This would also help improve the performance of the existing models because of the necessity of more data for handling 18 strategies. Similarly, an approach that models this problem as a multi-label problem as opposed to independent binary classifiers would be made possible by more data.

Acknowledgments

We would like to thank Dr. Dan Dinsmore (Associate Dean of Research, External Funding and Faculty Development, University of North Florida) and Dr. Katherine Hopper (Associate Lecturer, Department of Psychology, University of North Florida) for providing the fNIRS data that was used in this study. The authors would also like to thank Dr. Tara Rowe (Director of THRIVE, University of North Florida) for the assistance in writing the paper.

References

[1] S. Burns. M. Lieberman. "The use of functional near infrared spectroscopy (fNIRS) for unique contributions to social and affective neuroscience". *PsyArXiv*, 2019.

[2] M.M. Plichta. M.J. Herrmann. C.G. Baehne. C. Ehlis. M.M. Richter, P. Pauli. A.J. Fallgatter. "Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?". *Neuroimage* ., 2006.

[3] J. Berrenadi. H. Maior. A. Marinescu. J. Clos. M.L. Wilson. "Exploring Machine Learning Approaches for Classifying Mental Workload using fNIRS Data from HCI Tasks". *Journal of Human Ergology*, 2019.

[4] Y. Zhang. Y. Owechko. J. Zhang. "Driver cognitive workload estimation: A data-driven perspective". *The Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*, 2004.

[5] A. Girouard. E.T. Solovey. R.J.K. Jacob. "Designing a passive brain-computer interface using real time classification of functional near-infrared spectroscopy". *International Journal of Autonomous and Adaptive Communications Systems (IJAACS)*, vol.6, no.1, 2013.

[6] A. Sassaroli. F. Zheng. L.M. Hirshfield. A. Girouard. E.T. Solovey. S. Fantini. ”Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy”. *Journal of Innovative Optical Health Sciences*, vol.1, no.2, pp. 227-237, 2008.

[7] D. Afergan. E.M. Peck. E.T. Solovey. A. Jenkins. S.W. Hincks. E.T. Brown. R. Chang. R.J.K. Jacob. ”Dynamic difficulty using brain metrics of workload”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[8] A. Girouard. E.T. Solovey. L.M. Hirshfield. E.M. Peck. K. Chauncey. A. Sassaroli. S. Fantini. R.J.K. Jacob. ”From brain signals to adaptive interfaces: using fNIRS in HCI.”. *Brain-Computer Interfaces*, 2010.

[9] B.F. Yuksel. K.B. Oleson. L. Harrison. E.M. Peck. D. Afergan. R. Chang. R.J.K. Jacob. ”Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state”. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.