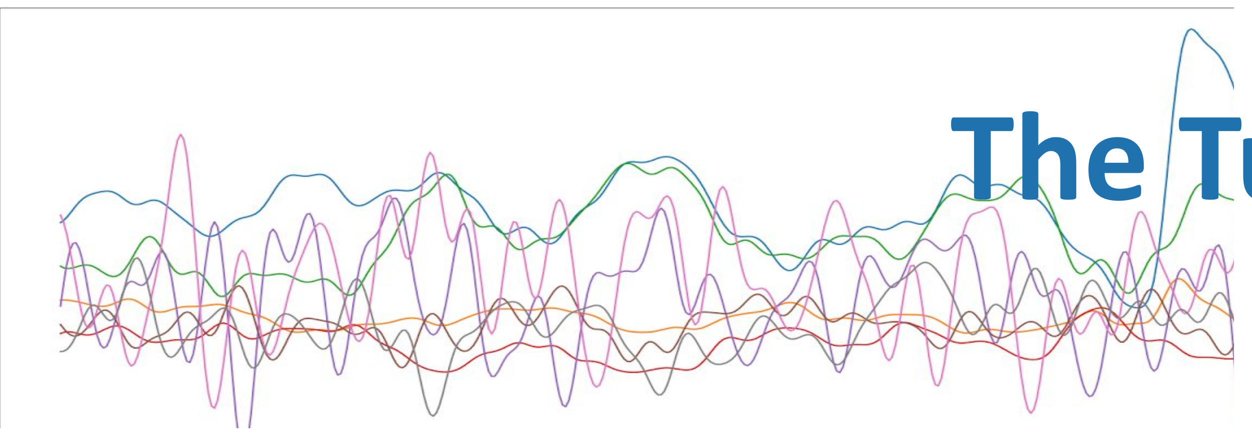


The Tufts fNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces that Generalize

Tufts UNIVERSITY

Zhe Huang*[1], Liang Wang*[1], Giles Blaney[2], Christopher Slaughter[3], Devon McKeon[1], Ziyu Zhou[1], Robert Jacob*[1], and Michael C Hughes*[1]

1 Dept. of Computer Science, Tufts University *2 Dept. of Biomedical Engineering, Tufts University* *3 Dept. of Computer Engineering, Univ. of Maryland Baltimore County*

Dataset: https://tufts-hci-lab.github.io/code_and_datasets/fNIRS2MW.html Code: https://github.com/tufts-ml/fNIRS-mental-workload-classifiers

Overview

Goal:

Enable everyday BCI by building classifiers that predict mental state using passive fNIRS reading (e.g., 30-second window)

Barriers:

1. Lack of large-scale open-access fNIRS dataset
2. Lack of standardized training and evaluation protocols

Contributions:

1. *Released the largest open-access fNIRS dataset*
2. *Proposed standardized training and evaluation protocols*
3. *Provided extensive benchmark results*

Possible Future Uses

1. *Sliding-window time series classification*
2. *Domain generalization*
3. *Fairness in time series classification*

Dataset

68 subjects. Largest open-access fNIRS dataset

Time Series Classification Task

***Input*: short window of multivariate fNIRS**

***Output*: mental workload intensity level (low vs high)**

Features: fNIRS measurements

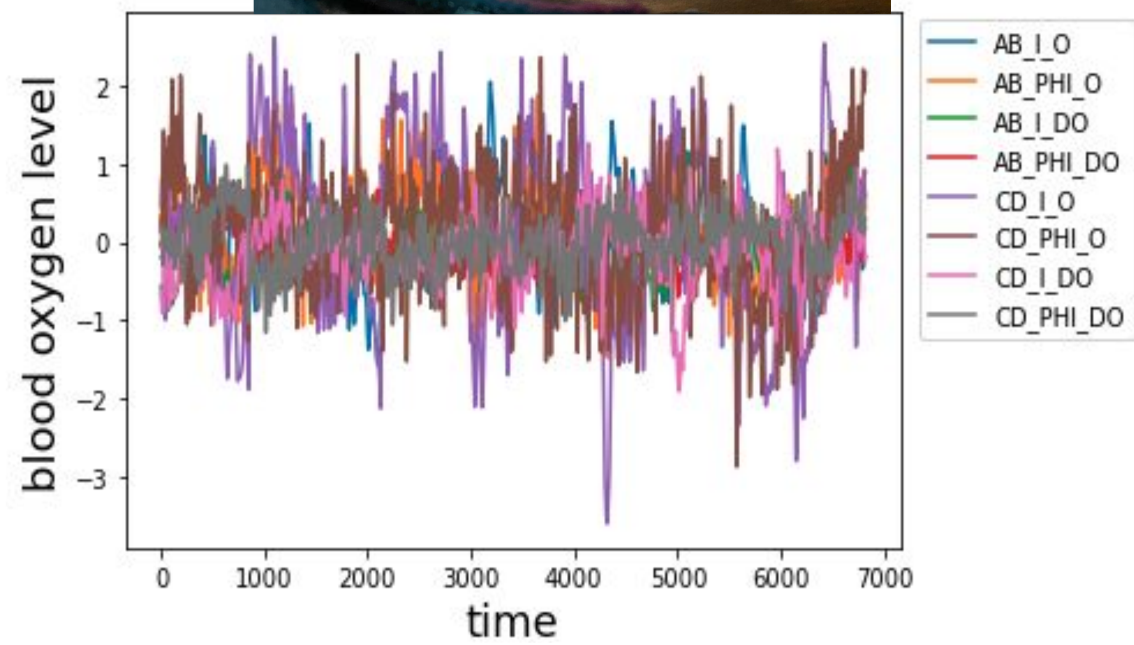
- 8 channels at 5.2 Hz
- Preprocessing to remove artifacts

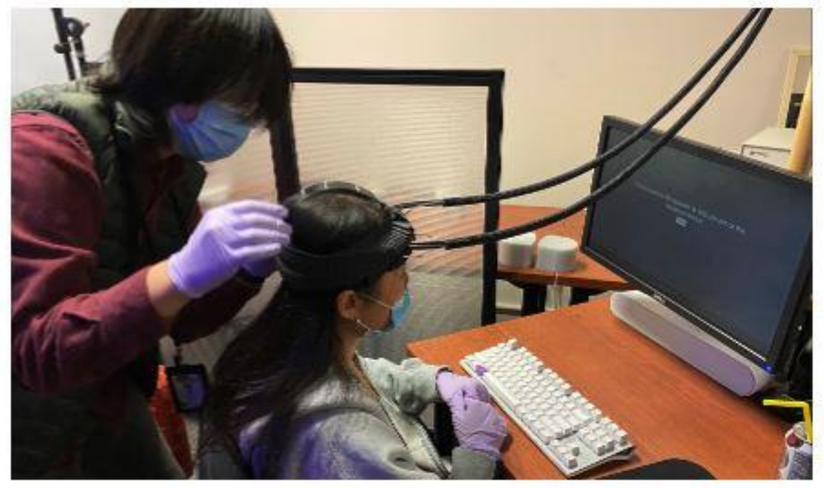
Labels : Mental workload intensity

- *n-back task:* standard experiment for measuring memory workload

Other metadata are also included:

Race, Sex, Age, Handedness etc



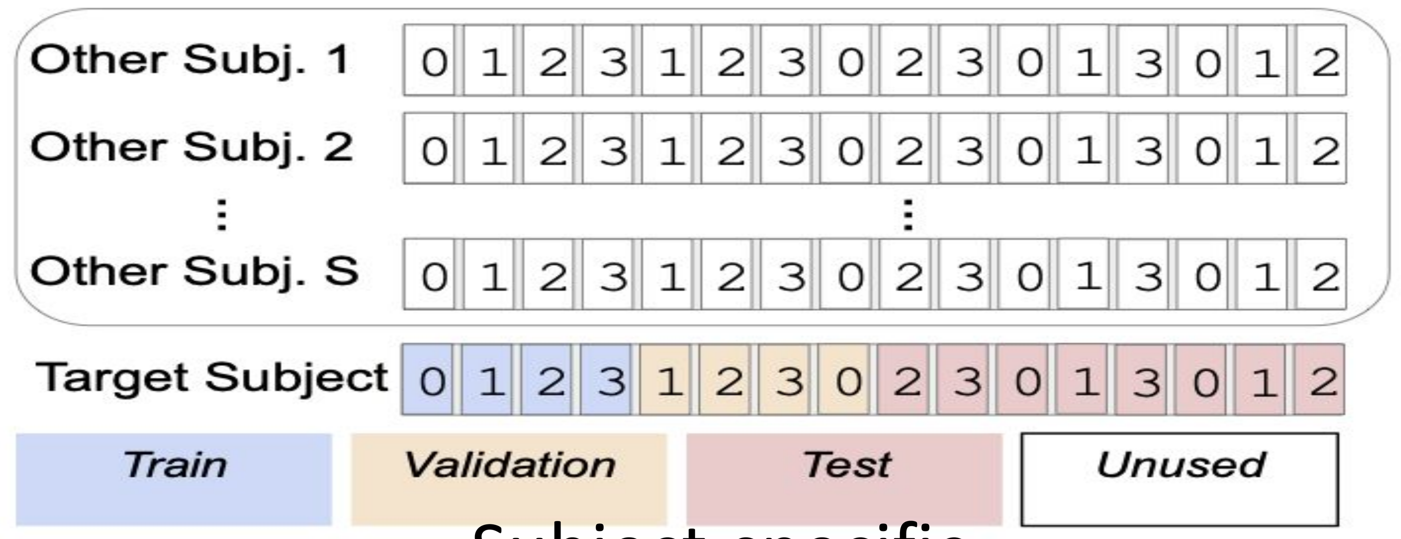


(a) A subject wearing the fNIRS headband, placed by the operator before each experiment began.

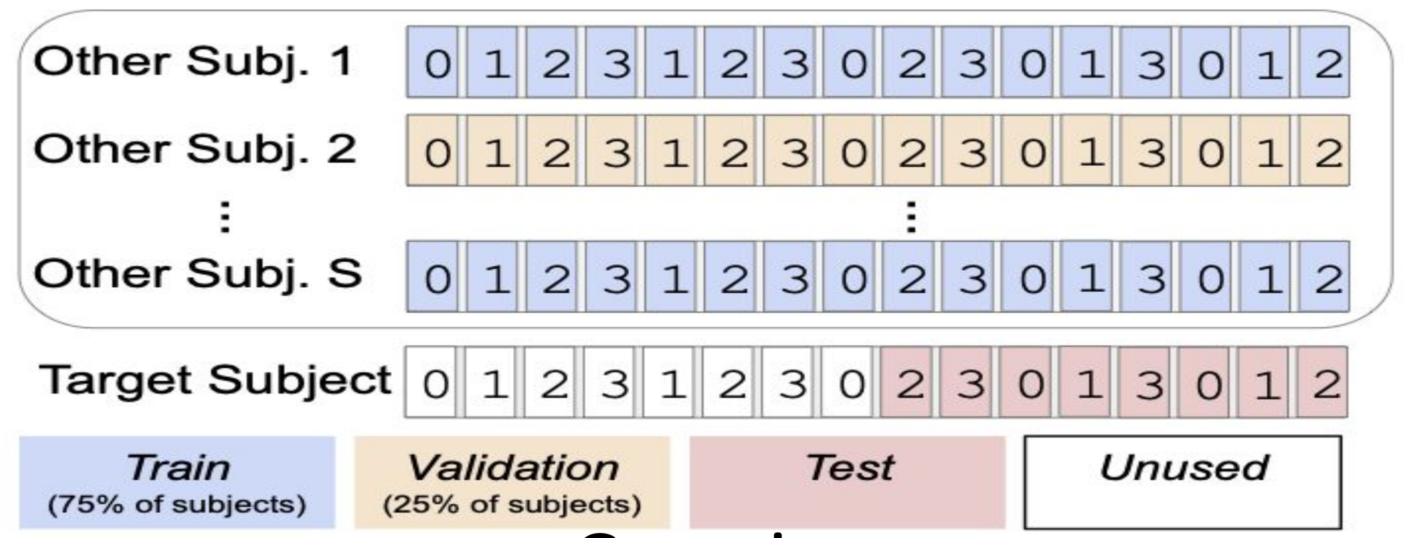
	Asian	White	Latin	Black	Pac. Isl.	other
race	32	27	3	2	1	3
	M	F	other	decline		
gender	26	39	2	1		
	right	left	unk.			
handedness	64	3	1			
	min.	max.	mean	std		
age	18.0	44.0	21.71	4.01		
sleep last night (hr.)	3.5	10.0	7.29	1.21		

(b) Demographics of our eligible cohort (n=68).

Training & Evaluation Protocols



Subject-specific



Generic

Generic + Fine tuning

Audit of Fairness across Subpopulations

Our large dataset with demographic labels enables audits of generalization across subgroups

Train Group	Test Group	RF Accuracy	EEGNet Accuracy
White (21)	White (6)	77.44 (66.70 - 87.31)	67.69 (54.11, 80.87)
	Asian (6)	74.18 (63.49 - 85.19)	64.88 (53.61, 76.54)
	URM (6)	71.43 (60.60 - 82.14)	62.41 (48.21, 76.31)
Asian (26)	Asian (6)	67.91 (59.01 - 77.56)	64.14 (53.81, 75.98)
	White (6)	71.29 (59.83 - 81.77)	64.98 (54.76, 75.01)
	URM (6)	67.39 (57.86 - 77.44)	65.22 (53.10, 77.43)

Benchmark Results

1. Generic classifiers benefit noticeably from larger training set
2. Generic classifiers outperform both subject-specific and fine-tuning classifiers
3. Substantial variation exists across users