TUFT Dataset – Detailed File Report

# 1. fnirs\_init.csv

This file contains the initial resting-state fNIRS signals collected from the subject before any task was introduced. It acts as a reference dataset that captures the subject’s baseline brain activity under neutral or non-task conditions. The data is stored in CSV format, representing a multichannel time-series, where each row corresponds to a time point and each column represents an individual fNIRS channel—measuring hemodynamic responses such as oxyhemoglobin (HbO) or deoxyhemoglobin (HbR) concentrations. This baseline signal is critical for ensuring that any changes observed during task sessions can be attributed to cognitive activity rather than individual variability or sensor drift.

* Purpose:
* Establishes baseline brain oxygenation levels prior to task performance.
* Helps normalize fnirs\_train.csv and fnirs\_test.csv to correct for signal drift and inter-subject variability.
* Useful in preprocessing pipelines to standardize signals and improve comparability across sessions and subjects.

# 2. fnirs\_train.csv

This file contains the time-series fNIRS data collected while the subject was actively performing cognitive tasks during the training session. The structure of this file mirrors that of fnirs\_init.csv, with each row representing a time point and each column representing a distinct measurement channel. These recordings capture dynamic brain responses to the known stimuli presented during the training session, the details of which are logged in the corresponding task result files. This file serves as the main physiological input for supervised learning and neurocognitive analysis.

* Purpose:
* Provides multichannel fNIRS signal data during labeled training tasks.
* Can be precisely aligned with task\_result\_train\_\*.txt files to extract task-specific signal segments.
* Enables training of models to classify brain states, cognitive workload levels, or predict behavioral performance.

# 3. fnirs\_test.csv

This file contains the fNIRS signal data recorded during the test session, in which tasks are presented without explicit feedback or label guidance at the time of testing. The structure is identical to the training data file, enabling a straightforward comparison between the two. The signals recorded here are used to evaluate model performance on unseen data, and they can also be analyzed for understanding brain activity patterns during more complex or unfamiliar tasks.

* Purpose:
* Serves as the physiological dataset for the testing phase.
* Can be used to validate or test models trained using fnirs\_train.csv.
* Allows comparison of brain responses across varying task difficulties or cognitive load levels.

# 4. fnirs\_train.mat and fnirs\_test.mat

These files are MATLAB-formatted binary versions of the training and testing fNIRS signal data. While they include the same raw signal values found in the .csv files, they often additionally encapsulate structured metadata such as sampling rates, channel names, and timestamps. This additional context can be particularly beneficial when performing advanced analyses using MATLAB-based tools. These files are compatible with standard fNIRS toolboxes and streamline integration into neuroimaging workflows.

* Purpose:
* Enables fNIRS signal analysis using MATLAB or domain-specific libraries like HOMER2 or NIRS-SPM.
* Contains structured metadata that aids in reproducibility and advanced signal interpretation.
* Facilitates extended analyses such as brain connectivity mapping, time-frequency decomposition, or channel-specific filtering.

# 5. task\_result\_train\_1\_1\_1.txt

This text file captures the behavioral performance data from a training session (identified as session 1\_1\_1). It logs the sequence of presented stimuli, the participant’s responses via the keydownSpaceSequence, the correct responses in gold\_standard, and the resulting accuracy. Additionally, the bigN field indicates the cognitive load or complexity of the task. In this particular case, bigN = 0 suggests a simple task, and the perfect accuracy (1.0) reflects flawless performance by the subject.

* Purpose:
* Records detailed behavioral response data during training tasks.
* Links each stimulus and response with corresponding fNIRS signals in fnirs\_train.csv.
* Serves as a label source for supervised learning—useful for classifying correct vs. incorrect trials.
* Provides the bigN value for identifying cognitive load levels associated with each session.

# 6. task\_result\_test\_1\_2\_1.txt

This file documents the subject’s performance during a test session (1\_2\_1) under higher cognitive load (bigN = 3). Similar to the training task result file, it includes the stimulus sequence, participant responses, ground truth answers, and an overall accuracy score (in this case, 0.825). The reduced accuracy compared to training suggests increased difficulty or cognitive challenge, making it useful for evaluating how well models or analyses generalize to more demanding conditions.

* Purpose:
* Captures behavioral outcomes for test sessions under higher task complexity.
* Provides validation targets for models built using training data.
* Enables analysis of how fNIRS signals correlate with behavioral metrics under cognitive strain.
* Supports load-based stratification of signal data for workload estimation.

# Relationships Between Files

The dataset’s structure establishes a clear mapping between physiological data (.csv, .mat) and behavioral data (.txt) for both training and testing sessions. Each session is identified by a serial code (e.g., 1\_1\_1, 1\_2\_1), ensuring traceability and alignment across modalities. This structure allows for synchronized analysis, where segments of the fNIRS signal can be directly associated with specific stimuli, responses, and performance metrics. Such integration is critical for conducting comprehensive analyses that span both neural and behavioral dimensions.

* Purpose:
* Allows synchronized alignment of time-series brain signals with behavioral events and outcomes.
* Enables development of supervised and unsupervised machine learning models linking physiology to cognition.
* Supports session-wise and condition-based segmentation of data for targeted analysis and interpretation.