**Skyn Data Manager (SDM)**

**Instructional Manual**

Status… *Experimental testing*

Version… 0.0.1

Last updated… *7.21.2023*

For help… *email* [*ndidier@bsd.uchicago.edu*](mailto:ndidier@bsd.uchicago.edu)

**Table of Contents**

Page 2… Software Capabilities

Page 3… Software Setup

Page 4… Data Setup

Page 5… Metadata Setup

Page 6 + 7… Running software

**SDM Functions:**

*Signal processing*

* Datasets can be cropped according to user-designated durations/times
* Removes artifacts and smooths the TAC signal for each dataset
* Provides temperature visualization and detection of potential non-wear

*Feature engineering*

* For each Skyn dataset, the SDM will calculate: Peak, TAC-AUC, Rise Duration, Fall Duration, Rise Rate, Fall Rate, and several others features.

*Make Predictions on New Data using Built-In Models*

* A random forest model and logistic regression model can be used to make predictions of whether or not a single Skyn dataset corresponds to alcohol consumption
* Model was trained & tested using group k-fold cross validation with 30 alcohol episodes and 30 non-alcohol episodes.
* Random forest and logistic regression achieved accuracy of 97% (58/60)

*Train and Make Predictions with New Models using New Data*

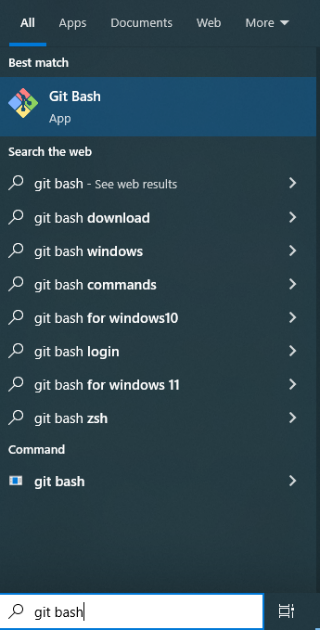
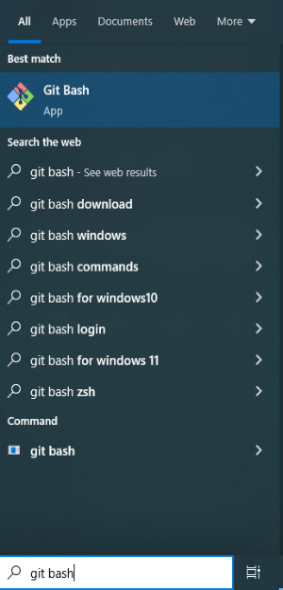
* 2 models are trained: random forest and logistic regression
* Models are trained in two ways: using features from cleaned (processed) data and using features from raw (unprocessed) data
* Altogether, 4 models will be trained and tested

*Automated Excel Workbooks and Reports*

* Each individual dataset will saved as an Excel file that includes 3 tabs:
  + Processed Skyn datasets will be saved with new columns for cleaned & smoothed data
  + Variable key that defines the column names
  + Graphs to visualize the raw versus cleaned data, temperature, etc.
* Whole-cohort analysis will produce an Excel file that includes several tabs for:
  + Feature data
  + Model prediction results
  + Compiled visual summary of each Skyn dataset
* Several graphs summarizing the models

**Software Installations and Setup:**

1. Download git: [Git - Downloads (git-scm.com)](https://git-scm.com/downloads)
   * This makes it easy to download the SDM from Github
2. Download Anaconda: [Anaconda | The World’s Most Popular Data Science Platform](https://www.anaconda.com/)



* + This is a software package that includes Python so SDM can run. It also includes several IDEs such as VS Code, if you would like to open and edit the code.

1. Open git bash (if on Windows) or terminal (if on Mac)
2. Navigate to the folder you would like to store the SDM:
   * Text

     Description automatically generatedcd <path to folder>
3. Download SDM:
   * A screenshot of a computer screen

     Description automatically generatedgit clone <https://github.com/ndidier3/skyn_data_manager.git>

Folder should look like this:

**Data Setup:**

*Cohort* includes each Skyn dataset that you would like to analyze, usually files associated with a set of participants who completed the same methodological paradigm.

Note: failing to follow the below guidelines will cause software failure*.*

* All cohort data must be **contained within the same folder**
* Each file must be a **.csv** or **.xlsx**
* Filenames must have a **consistent naming convention**:
  + Example filename: #5001 ^Alc $001.csv
    - **SubID**: 5001
    - **Condition**: Alc
    - **Sub-condition**: 001
  + SubID should be used to identify each participant.
  + Condition should either be ‘Alc’ or ‘Non’ to indicate whether alcohol was consumed (based on self-report or some other measure)
    - For testing new data on out-of-the-box predictive models, Condition can be ‘Unknown’. However, this will prevent the generation of performance metrics (accuracy, etc.).
  + A screenshot of a computer

    Description automatically generated with medium confidenceSubID and Condition are always necessary.
  + Sub-condition is not necessary unless the same subject has more than 1 alcohol dataset or more than 1 non-alcohol dataset.
    - The Sub-condition differentiates these datasets
    - If Sub-condition is needed, it must be written for each filename.
  + SubID, Condition, Sub-condition should be preceded by a distinct character, such as # or $

Example of naming convention. Notice the consistent spacing and use of #, $, and ^

* + - These characters will tell the software later how to automatically identify SubID, Condition, and Sub-Condition

**Metadata Setup:**

* Metadata files are stored in the resources/ folder

*Text

Description automatically generated*

* Do not edit the *MARS\_original\_testing* or *Test* folders.
  + These folders include metadata for running the software on the original MARS dataset (Didier et al., under review) and a fake testing dataset.
* Do not edit the FeatureKey spreadsheet.
  + However, you can view it to review feature descriptions.
  + This key will be included in cohort report documents.
* *A screenshot of a computer

  Description automatically generated with low confidence*Fill in the Excel document titled “Cohort Metadata.”
  + For each Skyn dataset, add a row that labels the SubID, Condition, and Sub-condition for that dataset.
  + If a file should be left out of analyses, you can exclude with “N” in the Use\_Data column.
  + Optional: Fill in the TotalDrks column to indicate the total drinks consumed during the drinking episode
* Optional: If you have timestamps to indicate the beginning of drinking episodes, fill in the Excel document titled “Timestamps Metadata”
  + Format requirements for dates and times:
    - Dates: day/month/year (25/5/2023) or month/day/year (5/25/2023)
    - Use military time
    - Optional: Time zones
      * Table

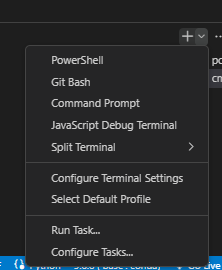
        Description automatically generatedTime zones will be used to standardize Skyn-generated timestamps.
      * Time zones reported here must **reflect** **where the user wore the Skyn**.
      * If Skyn data is downloaded in a different time zone from the user, then timestamps in the Skyn dataset will not reflect the user’s local time.
      * If time zone is provided, the software will adjust the timestamps in the Skyn dataset to ensure it matches the Skyn user’s local time.
      * If time zone is not provided, software will assume that the time zone of Skyn data and time zone of the Skyn user are equivalent.

A picture containing text

Description automatically generated**How to Run Software:**

* Open VS Code
* File -> Open Folder, and open “skyn\_data\_manager”
* On left panel, click on “code” folder, then click on “params.py”
  + If you are asked to select an interpreter, choose “Anaconda3/python.exe” or similar Python3 interpreter
  + A screenshot of a computer

    Description automatically generated with medium confidenceModify “cohort\_name” to a name of your choosing
  + Modify “data\_in” to let the software know the location of your cohort data folder
  + If needed, modify the other parameters listed
    - The default search\_character and search\_length parameters match the filenaming convention shown to the right.
    - If you do not use sub\_conditions, set
      * sub\_condition\_search\_character = None
      * sub\_condition\_search\_length = None



For Windows users only (above). Mac users will use terminal app.

After params.py is updated:

* WINDOWS ONLY:
  + In VS Code, Terminal -> New Terminal
  + Click the + Icon, then click Command Prompt
  + Ideally, two commands are automatically run:
    - C:/Users/<user> /Anaconda3/Scripts/activate
    - *conda activate base*
  + If not, run command:
    - *conda activate base*
* MAC ONLY:
  + open Mac terminal (this can be found in Applications)
  + run terminal command: *conda activate*
* Then run command:
  + *pip install -r code/requirements.txt*

**Now you are ready to process your Skyn data.**

* To test software, run:
  + *python code/process\_and\_predict\_TEST.py*
* To run signal processing, feature engineering, and make predictions using default built-in model on your cohort of data:
  + *python code/process\_and\_predict.py*
* To run signal processing and feature engineering**, but build a new model to make predictions***:*
  + *Python code/process\_train\_and\_predict.py*
* See processed\_data\_and\_plots folder for individualized results
* See features\_and\_model\_results for machine learning and feature results for whole cohort