**Skyn Data Manager (SDM)**

**Instructional Manual**

Status… *Experimental testing*

Version… 0.0.1

Last updated… *10.23.2023*

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

**Table of Contents**

Page 2… Summary of Software Capabilities

Page 3… Initial Software Setup

Page 4-9… Opening and Running software

Page 10… Data Requirements

Page 11… Metadata Requirements

**SDM Functions:**

*File Configuration*

* Convenient file renaming tool to match SDM filename convention
* File splitting tool for converting multi-day datasets into separate Excel files

*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*

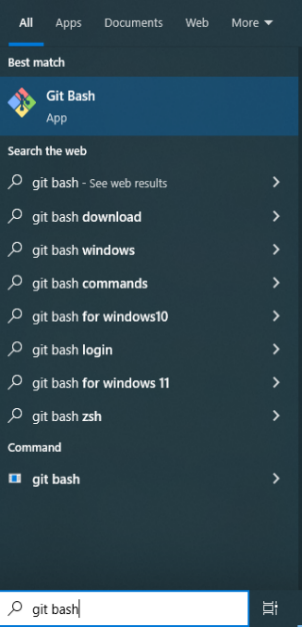
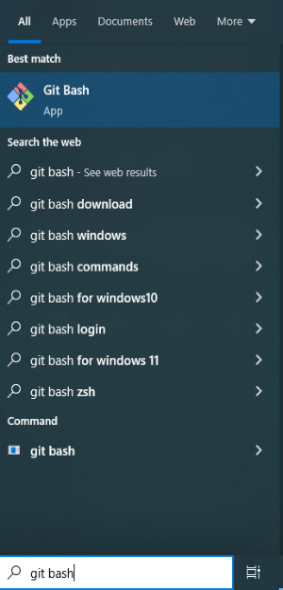
* Each individual dataset will saved as an Excel file that includes 3 tabs:
  + Processed Skyn datasets will be saved with new columns for cleaned & smoothed data
  + Variable key that defines the column names
  + Graphs to visualize the raw versus cleaned data, temperature, etc.
* Whole-cohort analysis 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:

* If on Windows, download from here [Git - Downloads (git-scm.com)](https://git-scm.com/downloads)
* If on Mac, enter these 3 commands into the terminal:
  + /bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
  + brew update
  + brew install git

1. 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 – these are helpful if you would like to view or 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>
4. Check that folder looks like this:

A screenshot of a computer

Description automatically generated

* + App/ is where the software code exists
  + Raw/ is where raw Skyn data are stored
  + Resources/ include metadata and other files to help configure data
  + Results/ is where output files will be saved

**How to Use Software**

**Step 1:** Launch the app

* A blue background with white text

  Description automatically generatedOpen Anaconda Prompt (Windows) or terminal (Mac)
* Go to SDM folder with command:
  + *cd <path-to-folder>*

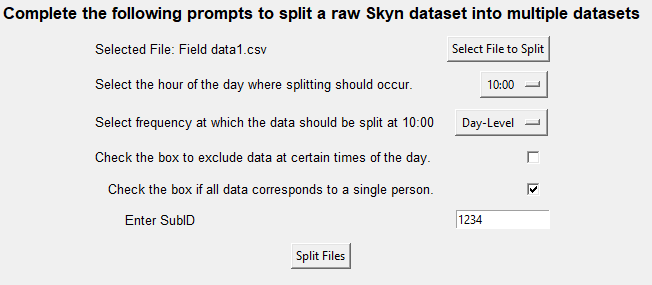
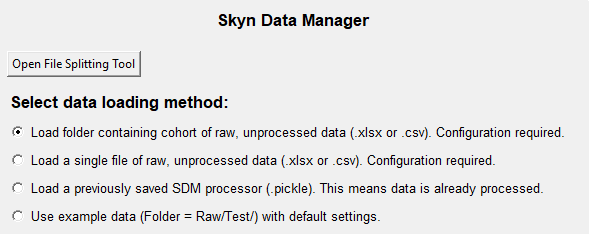
A black screen with white text

Description automatically generated

* **If first time using SDM,** run these below commands. Enter ‘y’ if commands have follow up questions. Some of these installs may take a few minutes.
  + *conda deactivate*
  + *conda create -n sdm-env python=3.8.8*
  + *conda activate sdm-env*
  + *conda install --file requirements.txt*
  + *conda install -c conda-forge kneed=0.7.0*
  + *conda install -c conda-forge scikit-learn=1.3.0* 
    - *this may take 10+ minutes, not because it takes up a lot of storage, but because Anaconda is creating compatibility between this package and all the other packages*



* **Launch SDM** using commands:
  + *conda activate sdm-env*
  + *python App/run\_sdm.py*



**Step 2**: Select data loading method.

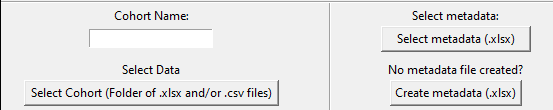
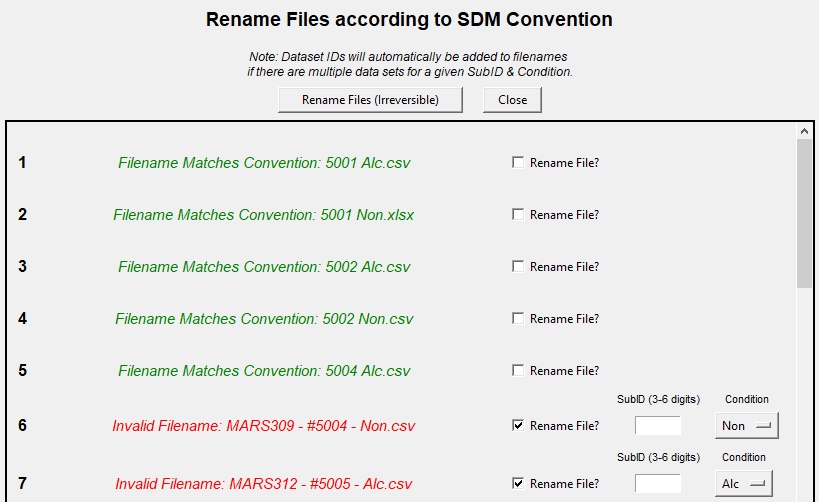
* If you have a single file containing multiple days of data and you want to make day-level predictions, this file must be split into separate data sets using the “**File Splitting Tool**”
* To conduct a quick test to make sure the software is setup and can be used on your computer, select the last option.
* To load a whole cohort of data, select the top option. After SDM processes a whole cohort, it will export the associated data and/or models within a .pickle file.
  + If you split a file using the File Splitting Tool, select the Split\_Files folder where all the separated data sets exist.
* To load only a single file, select the second option.
* The third option is for loading a .pickle file of an already-processed cohort.

**Step 3**: Select SDM program. Depending on the data loading method, SDM may display different settings or options.

A screenshot of a computer program

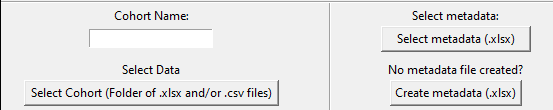
Description automatically generated

**Step 4.** Fill in Cohort Name and Select Data.



If filename(s) of the selected data are not compatible with SDM, you will be prompted to correct the filenames using SDM’s built-in file renaming tool.

**Step 5.** Select metadata if available. Otherwise, first use “**Create metadata**” tool.



**Step 6**. Modify Optional Settings (as needed)

**A screenshot of a computer program

Description automatically generated**

**Merging Files**

* If you have spreadsheets (demographics, other biomarkers, etc.) you would like to merge with SDM output, you can use “Merge Files with Results”.
  + This requires using SubID as a column key; SubID column must be available in each spreadsheet.
  + Details regarding requested merges will be displayed below

A screenshot of a computer

Description automatically generated

**Cropping Options**

* By default, SDM will not allow datasets to be longer than 24 hours. This can be changed in the “Crop Datasets” window.
* If a participant wears the device in a different time zone than the where the data was downloaded, then timestamps might be off by an hour or more. This may be a negligible concern, but if there is EMA-based timestamps (e.g. episode beginning) that you would like to match up with Skyn data, then this will ensure time zone alignment.
  + To adjust for this, you can provide the time zone (UTC) of where the data was uploaded. If participants uploaded their own data, enter 999
  + You must also upload a Timestamps spreadsheet so that SDM knows the time zone of each participant. This spreadsheet must also include episode start timestamps, so then SDM will crop data prior to these timestamps. See page 5, bottom for setting up Timestamps file.

**A screenshot of a computer screen

Description automatically generated**

**Loading Models**

* If your selected SDM program is to make predictions using already-trained models, you can select your own models to make these predictions. Predictions will be made on each dataset for each model provided, i.e., 4 models will result in 4 columns with either “correct” or “incorrect” for each dataset.
* This is not a requirement; SDM includes a built-in Random Forest model which should perform well in making binary predictions (Alc or no alc). Leave this option unchecked to use this default model.

**A screenshot of a computer program

Description automatically generated**

**Step 7.** When settings are all ready,click **RUN PROGRAM.**

**Data & Filename Requirements:**

Note: failing to follow the naming convention will cause software failure*.* *Cohort* refers to the collection of all Skyn datasets that you would like to analyze, usually files associated with a set of participants who completed the same methodological paradigm.

* All cohort data must be contained within the same folder
* Place your cohort data folder inside the Raw/ directory
* Valid File Types: csv or xlsx
* Required columns: TAC, motion, temperature, device ID/serial, timestamps, and firmware. These are the default columns included when Skyn data is downloaded.
* *A screenshot of a computer program

  Description automatically generated*Required filename convention (see filenames in image to right):
  + Example filename: 5001 Alc 001.csv
    - SubID: 5001
    - Condition: Alc
    - Dataset ID: 001
  + To rename cohort filenames according to this convention, use the SDM “Rename Files” tool (see page 7, Step 3)
* SubID is a unique ID assigned to each participant.
* Condition should either be ‘Alc’ or ‘Non’ to indicate whether alcohol was consumed (based on self-report or some other ground truth measure).
  + Condition can be ‘Unk’ when condition is unknown. However, this will prevent the generation of performance metrics (accuracy, etc.).
* SubID and Condition are always necessary.
* Dataset ID is not necessary unless the same subject has more than 1 alcohol dataset or more than 1 non-alcohol dataset.
  + The dataset ID differentiates each dataset
  + If dataset ID is needed, it must be written for each filename.

*Text

Description automatically generated***Resources Folder**

*The Resources folder includes Excel files of Metadata, Timestamps, and Feature Descriptions.*

* A metadata file is required to run SDM.
  + A screenshot of a spreadsheet

    Description automatically generatedColumn headers must include SubID, Condition, Dataset ID, Use\_Data, and TotalDrks
  + This file can be created manually in Excel or you can use the SDM “**Create Metadata**” tool (see page 7)
  + To exclude a dataset from analyses, label the row with “N” in the Use\_Data column.
  + *Optional*: Fill in TotalDrks column to indicate the # of drinks consumed for that episode.
* *Optional*: If you have timestamps to indicate the beginning of drinking episodes, fill in the Excel document titled “Timestamps Metadata”
  + Table

    Description automatically generatedFormat requirements for dates and times :
    - Dates: day/month/year (25/5/2023) or month/day/year (5/25/2023)
    - Use military time (HH:MM:SS)
* Do not edit the *MARS\_original\_testing* or *Test* folders.
  + These folders include metadata for running the software on the original MARS dataset and a test 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.