#### Project title: The effect of sampling frequency on mortality prediction in patients with TBI

Name: Feiya Xiang

Date: 2022/5/20

Update: 2022/11/28

#### Background

Traumatic brain injury (TBI) is defined as the disruption of normal brain function caused by an insult (bump, blow, or jolt) to the head. Worldwide, more than 50 million people experience a TBI per year and is estimated that about half the world’s population will experience a TBI within their lifetime. TBI is the leading cause of mortality in young adults and a major cause of death and disability across all ages in all countries.

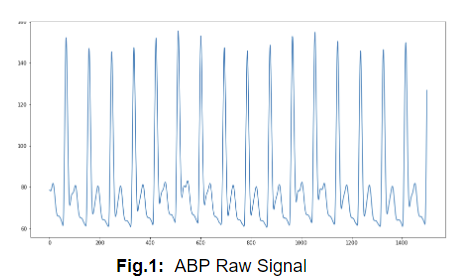
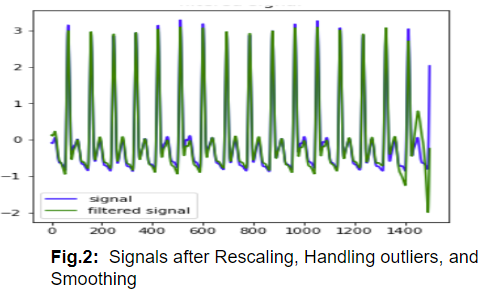
**Significance and Innovation**

Physiological signals monitoring is a cornerstone of neurocritical care after a traumatic brain. With the rapid development of artificial intelligence (AI) approaches to data analysis, the acquisition, storage, real-time analysis, and interpretation of physiological signal data can bring insights to the field of precision medicine. We review the existing literature on the quantification and analysis of the physiological signals and incorporate signal processing tools, advanced statistical methods, and deep learning techniques in order to comprehensively understand how the frequency of physiological signals affects TBI prediction accuracy.

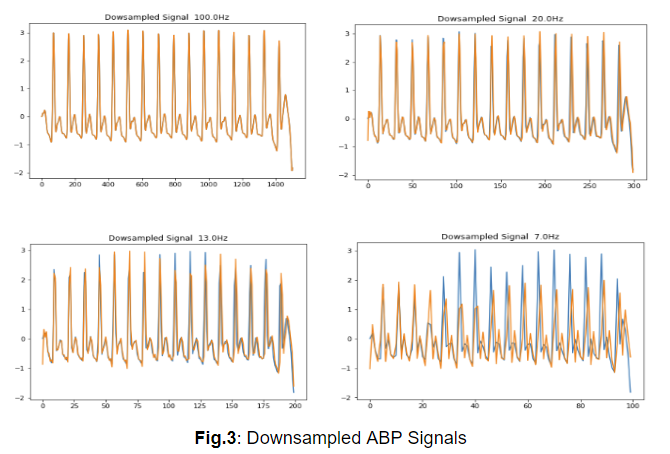
We will be using data from Center TBI databases to construct deep learning models of varying signal sampling frequency to predict the in-hospital mortality of patients hospitalized for TBI. Specifically, we’ll be focusing on the following signals: ECG, ABP, and ICP. First, we will extract all patient admitted to the ICU with a TBI diagnoses (based on ICD9/10 codes). The time series signals (discussed above) need to be filtered, cleaned, and segmented for the purposes of our sampling rate study. The ultimate goal of this project is to find best sampling frequency of these physiological time series signals which allows deep learning models as well as non-DL models to predict the clinical outcomes of patients with TBI. To achieve these objectives, we have broken down our project into following three aims.

**Aim 1: Curation of dataset (preprocessing of the data)**

I created a code repository to curate (handling outliers, signal processing, filtering, segmentation) and down sample the data. Time series signals measured in the real world is frequently non-stationary and riddled with noise/erroneous data, I utilized filters and signal smoothing techniques as a primary preprocessing step. Signal processing was achieved in prior studies via wavelet transform [1], moving average (MA) filter [2], and/or Savitzky-Golay filter [3].

Then, down sampling the data was done with 2 different means**: Interpolation** method(blue line) and **Fourier Transform** method(orange line). Downsampled 3 signals from 100 Hz to 1hz. All of these signals lose a lot of information from 20 Hz to 7 Hz, which can be shown as follows:

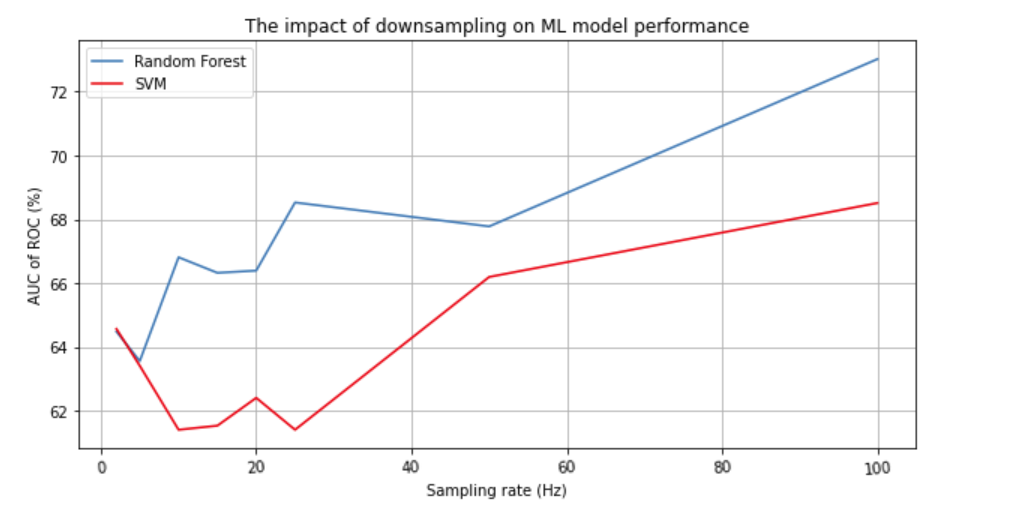


**Aim 2: Implementation of non-DL ML models**

As raw signals cannot be used for traditional non-DL ml models, some time has been spent curating features to be used in non-DL ML modeling. Before training, I have already extracted 9 **statistical features:** maximum, minimum, mean, median, mode, variance, range, kurtosis, skewness.

|  |  |  |
| --- | --- | --- |
| **Hyperparameters Tuning Results (on the original data)** | | |
| Models | Best AUC | Best parameters |
| AdaBoost | 0.6823 | {'algorithm': 'SAMME', 'learning\_rate': 1.1, 'n\_estimators': 100} |
| Random  Forest | **0.7202** | {'bootstrap': True, 'max\_depth': 10, 'max\_features': 'auto', 'min\_samples\_leaf': 4, 'min\_samples\_split': 2, 'n\_estimators': 50} |
| SVM | **0.6951** | {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf', 'probability': True} |
| LogisticRegression | 0.6158 | {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'} |

I have trained RF, SVM models on 100Hz, 50Hz, 25Hz, 20Hz, 15Hz, 10Hz, 5Hz and 2Hz signal datas. Their performance can be shown as follows:



**Current Roadblocks:**

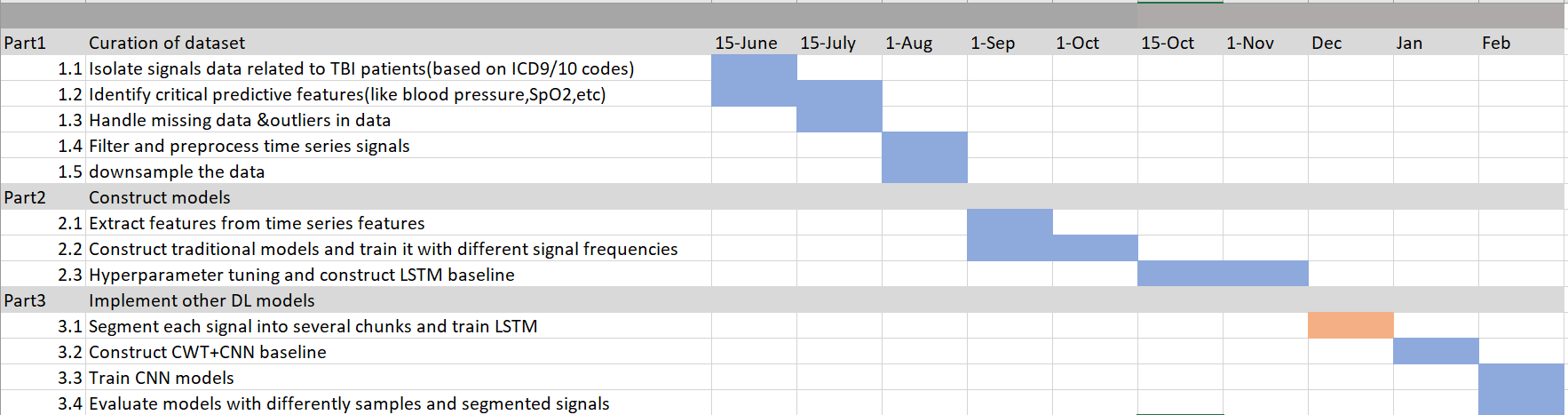
1. x-axis, that is sampling rate, should be changed to 100Hz, 90 Hz, 80Hz, 70Hz, 60Hz, 50Hz, 20Hz, 15Hz, 10Hz, 5Hz and 2Hz so that the relationship between sampling rate and model performance can be seen more clearly.
2. Even with a sampling frequency of 2hz, the model's accuracy score doesn't decrease that much. However, according to the downsampling procedure figures we drew earlier, these signals lose a lot of signal-based information when they are downsampled to 7hz. The reason may be that the **features we extracted** for traditional machine learning cannot effectively reflect signal-based information, that is, even when the fluctuation and shape of the signal have changed, these features have not changed much.

**Aim 3: Implementation of RNN, LSTM, and other DL models**

I have already segmented each signal into several chunks and constructed LSTM baseline. The next step is to train the LSTM model with different sampling rate signals.

Aside from LSTMs, using the Continuous Wavelet Transform and CNN to classify signals can be tried in the future since CWT can extract more wavelet features.

#### Timeline and Next Steps



**Citations**

[1] Leung, Howan, et al. "Wavelet-denoising of electroencephalogram and the absolute slope method: a new tool to improve electroencephalographic localization and lateralization." Clinical neurophysiology 120.7 (2009): 1273-1281.

[2] Guiñón, José Luis, et al. "Moving average and Savitzki-Golay smoothing filters using Mathcad." Papers ICEE 2007 (2007): 1-4.

[3] Press, William H., and Saul A. Teukolsky. "Savitzky‐Golay smoothing filters." Computers in Physics 4.6 (1990): 669-672.