DeepPurple

PI: Vahid Tarokh

Yu Xiang, Yuting Sun, Enmao Diao, Jie Ding, Jiawen Zhou

School of Eng. & Applied Sciences (SEAS)

Harvard University

Duke Data: Overview

- Main challenge: lack of data
    - Due to missing data***, only 18 subjects*** can be studied
    - Before inoculation, ***only two days*** of data are available

Duke Data

- **Goal: Predict infection after the inoculation within shortest time range**
- Focus on*: Physiological sensor data*
- Physiological sensor data (wearable device):
  - Electro-dermal activity (EDA),
  - Heart rate (HR)
  - Skin temperature (Temp)
- Features of physiological sensor data
  - Easy to collect and low expense
  - Less informative

Our Approach

- Our approach:

  (1) Apply Random Forest (RF) to make prediction

  (2) Use the prediction error as the features

- Key consideration:

  The model should make use of 24-hour data (one period) to make inference

Two Approaches

- Propose two approaches based on RF
  - The first approach serves as a preliminary model, which shows some promising features but is not quite stable
  - The second approach is an improvement of the first approach to make the model stable

First Approach

- Preprocessing of the data:
  - For each parameter (HR, TEMP, EDA), down-sample all the points in each hour to one point (take the median)
    1. 48-hour before inoculation: **a**=$\{a_1, \ldots, a_{48}\}$
    2. 24-hour after inoculation: **b**=$\{b_1, \ldots, b_{24}\}$
- Overview of the first method
  - Use **a**=$\{a_1, \ldots, a_{48}\}$ to train *one random forest* and make predictions $\hat{\boldsymbol{b}} = \{\hat{b}_1, \ldots, \hat{b}_{24}\}$
  - Distance(**b**, $\hat{\mathbf{b}}$) will be the features

First Approach: Details

- Initial training based on the 48 hours before inoculation + update the RF during prediction
  - Let $\boldsymbol{a}_k$=$\{a_k, a_{k+1}, \dots, a_{k+23}\}$ *to predict* a_{k+24}, $for\ k \in \{1, \dots, 24\}$
  - In words, use {k,…, k+23}-th hours to predict the (k+24)-th hour for $k \in \{1, \dots, 24\}$
  - Summary: one RF each with 24 variables and 24 training sets (con: training set too small so result not stable)
- Prediction: predict the 24 hours after inoculation
  - Predict the t-th hour using the trained random forest and the predicted data
  - Distance($\mathbf{b}$, $\hat{\mathbf{b}}$) will be the features: dynamic time warping (DTW)

Issues with the First Approach

- Training set too small compared with the number of features

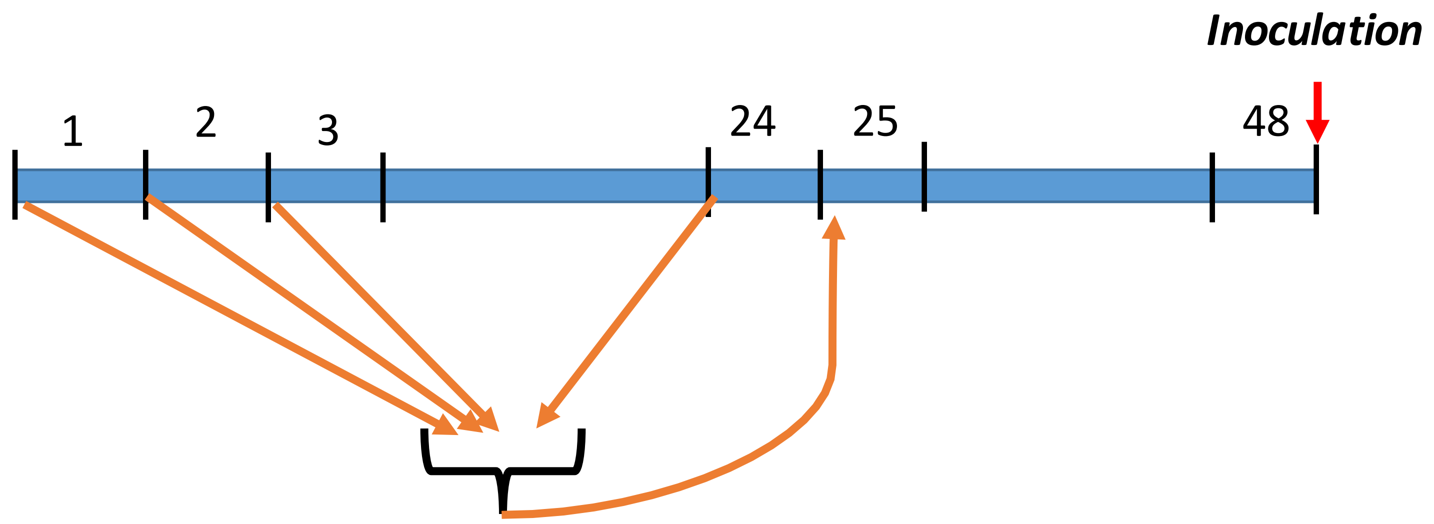
- Main improvement:
  - Increase training set by making better use of the data (down-sample to 1 point per minute)
  - Take into account the 24 hour periodicity

Second Approach

- Preprocessing of the data:
  - For each parameter (HR, TEMP, EDA), down-sample all the points in each minute to one point (take the median)
    1. 48-hour before inoculation: $48 \times 60 = 2880$ points
       a=$\{a_1, \ldots, a_{2880}\}$
    2. 24-hour after inoculation: 24 points **b**=$\{b_1, \ldots, b_{24}\}$
- Overview of the second method
  - Use **a** to train *one random forests* and make predictions
    $\widehat{\boldsymbol{b}} = \{\widehat{b}_1, \ldots, \widehat{b}_{24}\}$
  - Distance($\mathbf{b}$, $\hat{\mathbf{b}}$) will be the features

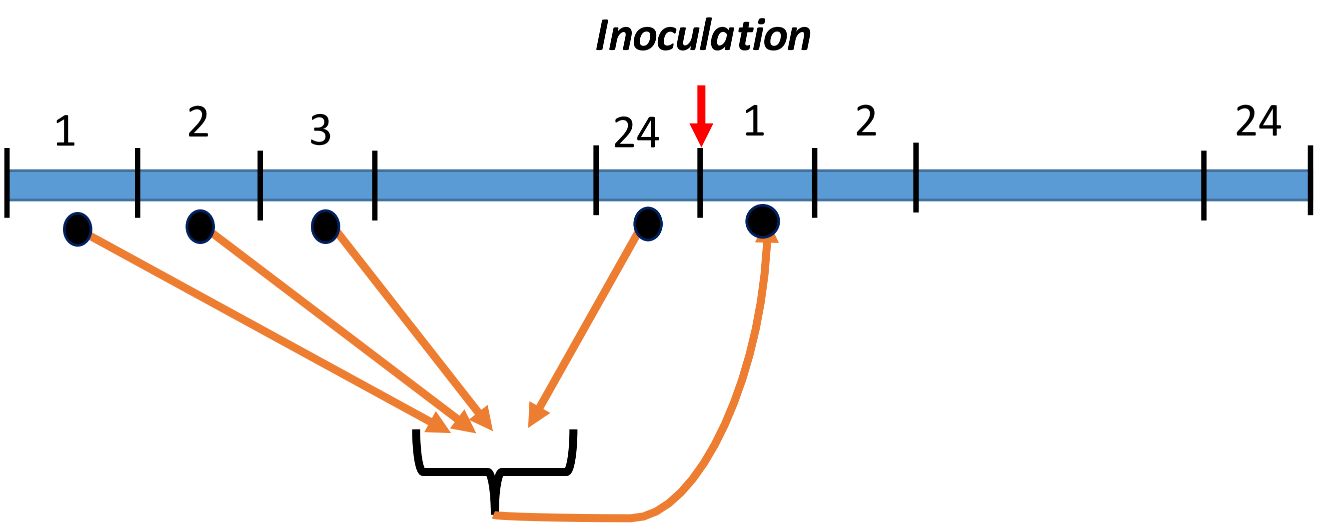
Second Approach: Details

- Initial training based on the 48 hours before inoculation + update the RF during prediction
  - Let $\boldsymbol{a}_k = \{a_k, a_{T+k}, a_{2T+k}, \ldots, a_{23T+k}\}$ *with* $T = 60$ *and* $k \in \{1, \ldots, 60\}$
  - In words, use the k-th minute of {1,…, 24}-th hours to predict the k-th minute of the 25-th hour and repeat this by sliding through all 48 hours
- Summary: one RF with 24 variables and about 60*48=2880 training sets



Second Approach: Details

- Prediction: predict the 24 hours after inoculation
  - First down-sample the 48 hours before inoculation to 48 points
  - Predict the t-th hour after inoculation using the random forest and the predicted data
  - Distance($\mathbf{b}$, $\hat{\mathbf{b}}$) will be the features: dynamic time warping (DTW)

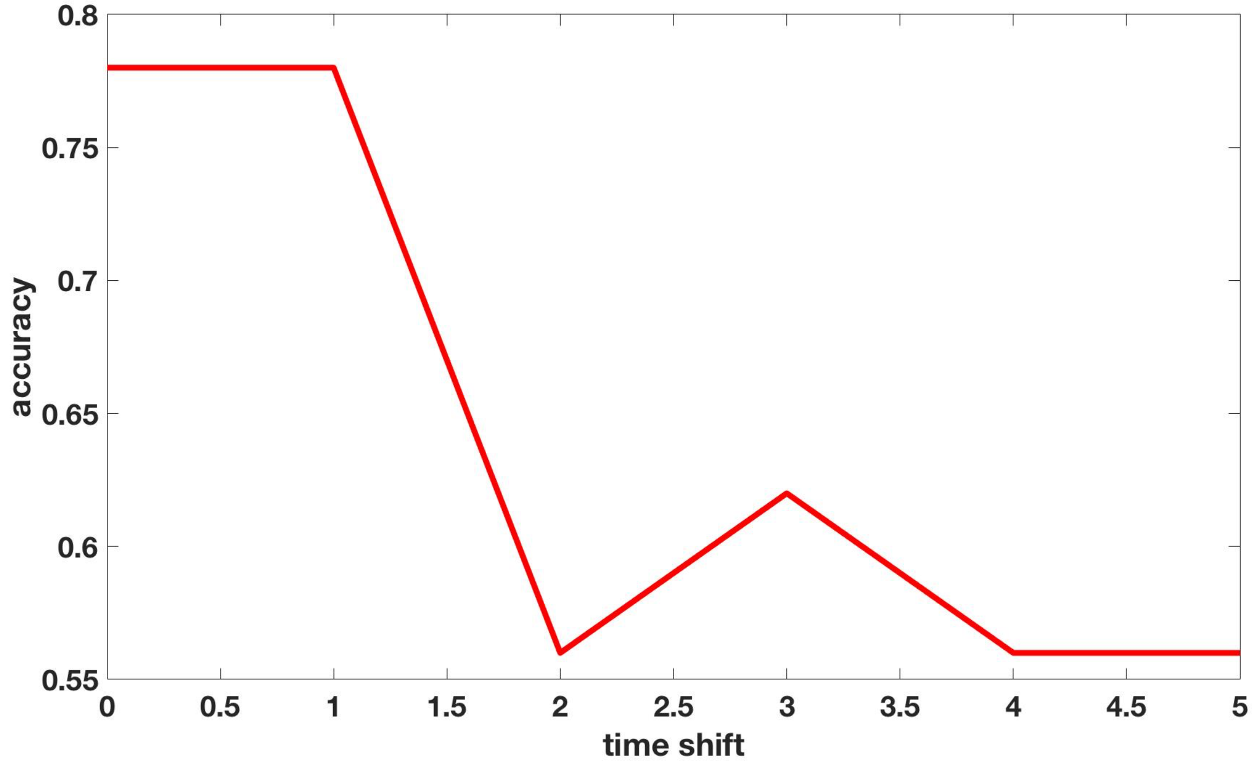


Summary Results

- True label (infected or not) is based on *symptom* and *shedding values*
    - *8 non-infected* subjects and *10 infected* subjects
- Use simple classification method: SVM
- Use cross validation to evaluation the performance

Summary Results

- Move the whole training and testing window (a 3-day window) *to the right* for 5 consecutive hours (step-size is 1 hour)
  - Did not move the window to the left due to lack of data



Summary Results

- Decompose two types of errors:
  - false positive rate = # false positive / total # of non-infected
  - false negative rate = # false negative / total # of infected