Predicting Clinical Outcomes with Wearables
Machine Learning from Small Data

Chenyang Lu
Cyber-Physical Systems Laboratory
Department of Computer Science & Engineering
https://www.cse.wustl.edu/~lu/
Machine Learning for Healthcare

- **Machine learning: data-driven tools for healthcare.**
  - Predicting clinical outcomes
  - Discover risk factors associated with outcome
  - Support clinical decisions to improve outcome

- **Discover the insights behind diverse data**
  - **Mobile Health (mHealth)**
    - Continuous monitoring outside hospitals
    - Small data: mHealth studies usually have moderate population
  - **Electronic Health Record (EHR)**
    - In-patient data
    - Big data: large patient population
Predict Readmissions with Fitbit

- Hospital readmission rate is high for heart failure patients.
  - ~25% patients readmitted within 30 days
- Predict deterioration (readmission + death) after discharge
  - Fitbit provides continuous monitoring of outpatients
  - Just-in-time intervention → better outcome and lower cost

Joint work with Thomas Bailey (Infectious Diseases), Marin Kollef (Critical Care), Dingwen Li (CSE)
Rich Features of Wearable Data

- Heart rate (HR), step count and sleep quality were collected.
  - Sampling period: 1 min (step, heart rate); 1 day (sleep)
- Statistical features:
  - First- and second-order features extracted from sliding window
  - 1st order: mean, max, min, skewness, kurtosis
  - 2nd order: energy, entropy, correlation, inertia and local homogeneity
- Detrended Fluctuation Analysis
  - Determine statistical self-affinity of time series
  - The fluctuation is then used as feature
- Sedentary behavior

![Sedentary behavior chart]

<table>
<thead>
<tr>
<th>sedentary bout</th>
<th>sedentary bout</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Important Features for Deterioration Early Warning
Predict Deterioration Risk

- **Input**: Fitbit data collected in the first N days since discharge
- **Deterioration**: composite outcome of either readmission or death
- **Small data**: 25 patients (18 with no deterioration vs. 7 deteriorated)
Assess Predictability

- Analysis of Variance (ANOVA) → predictability of outcomes
  - Test significant difference in the features between patients of different outcomes.
    - deteriorated patients vs. non-deteriorated patients
  - Assess feasibility of learning a stable predictor with the set of features.
- Significant differences in features → predictability

**Features**

<table>
<thead>
<tr>
<th>Features</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR skewness</td>
<td>7.9125</td>
<td>0.0099</td>
</tr>
<tr>
<td>HR correlation</td>
<td>5.4789</td>
<td>0.0283</td>
</tr>
<tr>
<td>HR DFA 10</td>
<td>5.3353</td>
<td>0.0302</td>
</tr>
<tr>
<td>Restless duration</td>
<td>5.2912</td>
<td>0.0308</td>
</tr>
<tr>
<td>Time in bed</td>
<td>5.2663</td>
<td>0.0312</td>
</tr>
</tbody>
</table>

Features with the largest F values.

*High F-statistic and low p-value → significant difference between group means.*
Feature Selection

- Select features using sequential forward feature selection
  - Avoid overfitting
  - Improve performance

- Features selected by the models have significant differences in ANOVA test

<table>
<thead>
<tr>
<th>Features</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep DFA 60</td>
<td>0.2254</td>
<td>0.6395</td>
</tr>
<tr>
<td>min asleep</td>
<td>4.3128</td>
<td>0.0492</td>
</tr>
<tr>
<td>daily step</td>
<td>4.3625</td>
<td>0.0480</td>
</tr>
<tr>
<td>restless count</td>
<td>4.2324</td>
<td>0.0512</td>
</tr>
<tr>
<td>awake count</td>
<td>4.0429</td>
<td>0.0562</td>
</tr>
<tr>
<td>min awake</td>
<td>2.2073</td>
<td>0.1509</td>
</tr>
<tr>
<td><strong>HR LH</strong></td>
<td>4.0282</td>
<td>0.0566</td>
</tr>
<tr>
<td><strong>HR DFA 10</strong></td>
<td>5.3353</td>
<td>0.0302</td>
</tr>
</tbody>
</table>
Test Overfitting

- Impact of the number of nearest neighbors (K) on KNN performance

**Smaller difference between training and testing errors \(\rightarrow\) less overfitting**
Evaluation with Small Data

- **Leave-one-out cross validation**
  - Leave a sample (or a group of samples) out for testing and train the model with rest of sample (or groups of samples).
  - Iterate through all the samples.

- For **imbalanced** dataset, do **not** just look at accuracy!
  - Example: for 1:9 positive/negative ratio, the predictor can achieve 0.9 accuracy if predicting everything as negative.
  - A good predictor should perform well for sensitivity, specificity and precision.

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{Specificity} = \frac{TN}{TN + FP} \quad \text{Precision} = \frac{TP}{TP + FP}
\]

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted positive</td>
<td>TP: true positives</td>
<td>FP: false positives</td>
</tr>
<tr>
<td>Predicted negative</td>
<td>FN: false negatives</td>
<td>TN: true negatives</td>
</tr>
</tbody>
</table>
**Deterioration Risk Prediction**

- Small data: 25 patients (18 with no deterioration vs. 7 with deterioration)
- KNN achieves higher specificity and precision than LACE index used in clinical practice.
- KNN performs better than neural network (NN)
  - Simple models help avoid overfitting on small data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>0.0770</td>
<td>0.9459</td>
<td>0.3333</td>
<td>0.720</td>
</tr>
<tr>
<td>KNN</td>
<td>0.5385</td>
<td><strong>0.9820</strong></td>
<td><strong>0.9130</strong></td>
<td><strong>0.8667</strong></td>
</tr>
<tr>
<td>LACE</td>
<td><strong>0.7647</strong></td>
<td>0.6250</td>
<td>0.5556</td>
<td>0.7826</td>
</tr>
</tbody>
</table>
Machine Learning with Small Data

- Apply statistical analysis to assess outcome predictability with the features.

- Mitigate overfitting in models
  - Adopt simple models
  - Reduce the number of variables through feature selection
  - Use an ensemble of models (Gradient Boosted Trees, AdaBoost)

- Handle imbalanced data sets
  - Random oversampling of the minority class(es)
  - Random undersampling of the majority class(es)
  - Synthetic minority over-sampling technique (SMOTE)

- Evaluation
  - Don’t just look at accuracy, especially for imbalanced dataset
  - Assess overfitting by comparing accuracy on training/testing data