Predicting Readmissions of Heart Failure Patients with Wearables

Dingwen Li
Cyber-Physical Systems Laboratory
Department of Computer Science & Engineering
Motivation

- Hospital readmission rate is high for heart failure patients.
  - ~25% patients readmitted within 30 days

- Early detection of deterioration (readmission + death)
  - Just-in-time intervention $\rightarrow$ better outcome
  - Reduce health care cost

- LACE index: state of practice for predicting readmission
  - Assess the risk of readmission based on inpatient data
  - Length (L) of stay; Acuity (A) of the admission; Co-morbidities (C); Number of emergency (E) department visits within the last 6 months

- Wearable devices provide convenient technologies for continuous outpatient monitoring.
Objective

- Assess the feasibility of collecting multi-modal data from wearable device.
- Explore the potential of predicting clinical deterioration
  - Develop and test predictive models based on machine learning
- Provide guidelines for future studies
Challenges

➢ Outpatient compliance
  - Wear devices continuously and properly outside hospitals
  - Potential problems
    • Forget to charge the device
    • Reluctant to wear the device
    • Improperly wear the device

➢ Learning from small sample size
  - May cause overfitting
  - Imbalanced dataset
Outline

- Study Setup
- System Analysis
- Predictive Model
System Overview

- FitBit
- Smart Phone
- Fitbit Cloud
- Transfer data
- Poll request
- HEROKU
25 heart failure patients were recruited.

- Each patient is given Fitbit ChargeHR wristband.
- Continuously monitor patients after discharge.
- Outcomes: 60-day deterioration events.

Heart rate (HR), step count and sleep status were collected.

- Sampling period: 1 min (step, heart rate); 1 day (sleep yield)
Outline

- Study Setup
- System Analysis
- Predictive Model
Data Collection Analysis

- **Yield**: fraction of the expected samples that are successfully collected and stored in our database.

- **Latency**: end-to-end latency from Fitbit to our database.
88% devices have step yield >0.8.
Only 60% devices have heart rate yield >0.8.
4 devices don’t have sleep data, 3 devices have low sleep yield.

Step monitoring has higher yield than heart rate and sleep
Compliance

Participants with yield <0.8 for all sensing modalities $\rightarrow$ non-compliant

92% participants generally followed the protocol
Potential Cause of Data Loss

- User compliance (not wearing the device) (2 patients)
  - Low yield for both heart rate and step

- Wearing device improperly (9 patients)
  - Large discrepancy between heart rate and step yield

- No connectivity between the wristband (0 patient)
  - Permanently losing un-sync data
End-to-End Latency

- Median: 8.6 minute; 99% percentile: 22.5 hour
- Did not cause data loss, as Fitbit can locally store data for 7 days.
- Latency allows daily intervention
- Latency can be reduced if analysis performed natively in Fitbit cloud
Outline

- Study Setup
- System Analysis
- Predictive Model
Data Preprocessing

- **Raw data format**
  - Raw heart rate and step count per minute
  - Derived data from Fitbit API, such as time in bed, sleep efficiency in sleep summary data

- **Data cleansing**
  - Extract step-related data only in 7am-7pm
  - Fill in missing data by carry forward imputation
  - Discard a whole day’s data if missing data last longer than the day
Feature Engineering

- **Statistical features:**
  - First- and second- order features extracted from sliding window
  - 1\textsuperscript{st} order: mean, max, min, skewness, kurtosis
  - 2\textsuperscript{nd} 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

- **Semantic features**
  - Sedentary behavior

---

10/10/19
Evaluation Methods

- **K-fold cross validation**: divide whole dataset into k equal-sized subgroups. Use one subgroup for testing and the remaining k-1 subgroups for training the model.

- **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).

Illustration of k-fold CV. Leave-one-out CV is a special case of k-fold CV with k=n.

For imbalanced dataset, it’s insufficient to 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 for imbalanced data should perform well for all of sensitivity, specificity and precision.

For clinical event prediction, we often prefer the predictors that have high sensitivity while fixing specificity at a certain level.

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

<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 Early Warning

- **Task**: Predicts upcoming deterioration using data collected in a recent sliding time window

- **Dataset**: 438 days’ data as samples with 427 normal samples and 11 deterioration samples

- **Model**: Weighted One Class SVM
  - Semi-supervised learning for anomaly detection
  - Need only majority samples to train model
  - SVM is sensitive to outliers
  - Sample weights suppress outliers in training
Deterioration Early Warning

- Impact of \textit{predicting horizon} and \textit{window size}

\begin{align*}
\text{Sensitivity} &= \frac{TP}{TP + FN} \\
\text{Specificity} &= \frac{TN}{TN + FP} \\
\text{Precision} &= \frac{TP}{TP + FP}
\end{align*}

More accurate in predicting recent events

More recent window is more informative
Different Predictive Models

High impact features
- Sleep and HR
- DFA → predictive features

<table>
<thead>
<tr>
<th>Model</th>
<th>Leave-one-subject-out</th>
<th>Leave-one-later-day-out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>K-means</td>
<td>0.9091</td>
<td>0.7500</td>
</tr>
<tr>
<td>LOF</td>
<td>0.0909</td>
<td>0.8500</td>
</tr>
<tr>
<td>OCSVM</td>
<td>0.8182</td>
<td><strong>0.9500</strong></td>
</tr>
<tr>
<td>WOCSVM</td>
<td><strong>0.9091</strong></td>
<td><strong>0.9500</strong></td>
</tr>
</tbody>
</table>

LOF: local outlier factor
Deterioration Risk Prediction

- **Task**: identifies participants at high risks of deterioration using data collected from the beginning of monitoring
- **Dataset**: 25 patients as samples, including 18 with no deterioration vs. 7 with deterioration
- **Model**: ML models for clinical readmission prediction
  - Random forest, SVM, logistic regression, multi-level perceptron, K-nearest neighbor
Analysis of Variance (ANOVA)

- Test whether the differences among group means are significant
  - High F-statistic and low p-value → significant difference among the group means.

- Indicate the feasibility of learning a stable predictor.

- Assumptions
  - Independent observations
  - The variable follow a normal distribution in each group
  - The variances within all groups must be equal (needed only for very imbalanced groups)
ANOVA Results

- Groups: deteriorated patients vs. non-deteriorated patients
- ANOVA test reveals significant differences in some features → suggesting feasibility of predictive models

<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
Feature Selection

- Select features using sequential forward feature selection
- 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>HR LH</td>
<td>4.0282</td>
<td>0.0566</td>
</tr>
<tr>
<td>HR DFA 10</td>
<td>5.3353</td>
<td>0.0302</td>
</tr>
</tbody>
</table>
Avoid Overfitting

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

![Graph showing training error vs. testing error](image)

**Smaller difference between training and testing errors \(\rightarrow\) less overfitting**
Evaluate Predictive Models

- Use all sensing modalities (step, HR, sleep)
- Use first 20 days’ data as input
- Specificity fixed at 0.95 to ensure an acceptable false alarm rate

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.7434</td>
<td>0.5551</td>
<td>0.3077</td>
<td>0.9459</td>
<td>0.6667</td>
<td>0.780</td>
</tr>
<tr>
<td>SVM</td>
<td>0.5943</td>
<td>0.4016</td>
<td>0.1795</td>
<td>0.9459</td>
<td>0.5385</td>
<td>0.7467</td>
</tr>
<tr>
<td>LR</td>
<td>0.7829</td>
<td>0.6118</td>
<td>0.4872</td>
<td>0.9009</td>
<td>0.6333</td>
<td>0.7933</td>
</tr>
<tr>
<td>NN</td>
<td>0.4002</td>
<td>0.2048</td>
<td>0.0770</td>
<td>0.9459</td>
<td>0.3333</td>
<td>0.720</td>
</tr>
<tr>
<td>KNN</td>
<td>0.7533</td>
<td>0.6680</td>
<td>0.5385</td>
<td>0.9820</td>
<td>0.9130</td>
<td>0.8667</td>
</tr>
<tr>
<td>LACE</td>
<td></td>
<td>0.7647</td>
<td>0.6250</td>
<td>0.5556</td>
<td>0.7826</td>
<td></td>
</tr>
</tbody>
</table>

AUC-ROC: area under Receiver Operating Characteristic curve
AUC-PR: area under precision-recall (sensitivity) curve

**KNN achieves higher specificity and precision than LACE index.**
Impacts of Multi-Modality

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>0.6910</td>
<td>0.5839</td>
<td>0.3333</td>
<td>0.9550</td>
<td>0.7222</td>
<td>0.7933</td>
</tr>
<tr>
<td>HR</td>
<td>0.8064</td>
<td>0.6502</td>
<td>0.3590</td>
<td>0.9369</td>
<td>0.6667</td>
<td>0.7867</td>
</tr>
<tr>
<td>Sleep</td>
<td>0.7237</td>
<td>0.3439</td>
<td>0.0513</td>
<td>0.9099</td>
<td>0.1667</td>
<td>0.6867</td>
</tr>
<tr>
<td>Sleep,HR</td>
<td>0.8064</td>
<td>0.6502</td>
<td>0.3590</td>
<td>0.9369</td>
<td>0.6667</td>
<td>0.7867</td>
</tr>
<tr>
<td>Sleep,Step</td>
<td>0.6556</td>
<td>0.6304</td>
<td>0.4615</td>
<td>1.0</td>
<td>1.0</td>
<td>0.860</td>
</tr>
<tr>
<td>Step,HR</td>
<td><strong>0.8151</strong></td>
<td><strong>0.6991</strong></td>
<td>0.4872</td>
<td>0.9369</td>
<td>0.7308</td>
<td>0.820</td>
</tr>
<tr>
<td>All</td>
<td>0.7533</td>
<td>0.6880</td>
<td><strong>0.5385</strong></td>
<td>0.9820</td>
<td>0.9130</td>
<td><strong>0.8667</strong></td>
</tr>
</tbody>
</table>

*Multiple sensing modalities improve predictive performance*
Number of Days Needed for Prediction

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-day</td>
<td>0.4247</td>
<td>0.2114</td>
<td>0.0256</td>
<td>0.9009</td>
<td>0.0833</td>
<td>0.6733</td>
</tr>
<tr>
<td>10-day</td>
<td>0.7339</td>
<td>0.6884</td>
<td>0.3077</td>
<td>1.0</td>
<td>1.0</td>
<td>0.820</td>
</tr>
<tr>
<td>15-day</td>
<td>0.7710</td>
<td>0.5808</td>
<td>0.3333</td>
<td>0.9009</td>
<td>0.5417</td>
<td>0.7533</td>
</tr>
<tr>
<td>20-day</td>
<td>0.7533</td>
<td>0.6880</td>
<td>0.5385</td>
<td>0.9820</td>
<td>0.9130</td>
<td>0.8667</td>
</tr>
</tbody>
</table>

Long-term monitoring (≥10 days) is needed for predicting readmission risks
Suggestions

- Try your best to get a large dataset

- If not, try statistical analysis first to assess feasibility

- Be careful with a small dataset
  - Don’t just look at accuracy, especially for imbalanced dataset
  - Assess overfitting by comparing errors with training/testing data sets
  - Applying techniques to mitigate overfitting

- Clean the dataset before training the model
  - Remove invalid values
  - Impute missing values
  - Normalize the dataset if needed