AI for Health with Wearables

Chenyang Lu
AI for Health Institute (AIHealth)
Wearables

- Commonly available: step, heart rate, sleep stages
- More sensing modalities
  - Oxygen saturation (SpO2)
  - Skin temperature
  - Breathing rate
  - Heart rate variability
  - ECG
  - Stress
- 500+ million wearables sold in 2021

Unobtrusive, longitudinal monitoring outside hospitals!
AI for Health with Wearables

- **Perioperative care**
  - Surgical complications after pancreatic surgery [Standard of Care in Surgical Prehabilitation]
  - Surgical outcomes of periacetabular osteotomy for hip dysplasia [NIH R01]
  - Recovery outcomes after spine surgery

- **Mental health care**
  - Depression and anxiety disorders [NIH All of Us]
  - Treatment response to behavioral therapy for depression and obesity [NIH R01]
  - Dynamic cognitive function in youth with diabetes
Wearable Data

- Time series: step count, heart rate, sleep stage

**Fine-grained, noisy, incomplete time-series**
Wearable Data → Clinical Outcomes

- **Reliable** prediction of clinical outcomes
  - Wearable data are fine-grained, noisy, incomplete

- **Scaling** predictive models
  - Small cohorts: robust models with limited data
  - Large population: deep learning models

- **Personalized** prediction of treatment response
  - Support personalized intervention

**Machine learning**

Predictions → Clinical decisions and intervention
Predict Clinical Outcomes with Wearables

- Complications after pancreatic surgery
  - Robust machine learning based on a small patient cohort

- Mental health disorders in the community
  - Deep learning for a large and diverse cohort

- Treatment response to depression therapy
  - Personalized prediction of treatment response
Need: Predict Outcomes of Pancreatic Surgery

- Pancreatic cancer has a 5-year survival rate less than 5%.

- Surgery is the only cure but commonly followed by complications.

- Predict postoperative complications before surgery
  - **Decision support:** suitability for surgery
  - **Intervention:** prehabilitation

Joint work with Chet Hammill (Surgery), Jingwen Zhang, Dingwen Li, Ruixuan Dai (CSE)
Goal: predict postoperative complications with preoperative data
- Wearable data collected by Fitbit wristband
- Patient clinical characteristics from medical record

Prospective study: 61 patients undergoing pancreatic surgery
- 25 (41%) experienced complications

Machine learning approach
- Small cohort → avoid deep models
- Features extracted from wearable data → shallow machine learning models
- Wearable data + Clinical characteristics
Robust Feature Engineering Pipeline

Daily Feature Extraction

Step Count
Heart Rate
Sleep Stage
Wearable Time Series

D: Number of Daily Features
N: Number of Days

High-level Feature Extraction

Inputs to predictive model

Clinical Characteristics

D N

Original Daily Feature
Extracted Trend

Day
Raw Time Series ➔ Daily Features

- **Statistical** features
  - max, sum, skewness, kurtosis, energy, entropy, inertia…

- **Semantic** features
  - Activity: daily steps, sedentary time, sedentary bout counts…
  - Sleep: awake counts, sleep efficiency, time of sleep, time before falling asleep…

- **Standardization**: account for missing data

---

```
Daily sedentary time

Duration of wearing the device

Daily sedentary time
#sample
```
Daily Features \(\rightarrow\) High-level Features

- Singular Spectrum Analysis (SSA)
  - **denoise** time series of daily features

![Trend Extraction with Complete Data (30 data samples)](image)
Daily Features → High-level Features

- Singular Spectrum Analysis (SSA)
  - denoise time series of daily features
Daily Features → High-level Features

- **Singular Spectrum Analysis (SSA)**
  - **denoise** time series of daily features.
  - **impute** missing daily features.

### Graphs

**Trend Extraction with Complete Data (30 data samples)**

**Trend Extraction with Missing Components (20 data samples)**
Daily Features → High-level Features

- Singular Spectrum Analysis (SSA)
  - denoise time series of daily features.
  - impute missing daily features.

Trend Extraction with Complete Data (30 data samples)

Trend Extraction with Missing Components (20 data samples)

- mean
- standard deviation
- slope
# Handling Missing and Noisy Data

## Standardization of Daily Features

<table>
<thead>
<tr>
<th>Daily Features</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-standardized</td>
<td>0.6919 (0.0116)</td>
<td>0.6332 (0.0289)</td>
<td>0.4320 (0.0560)</td>
<td>0.8611 (0.0000)</td>
</tr>
<tr>
<td>Standardized</td>
<td><strong>0.7326 (0.0074)</strong></td>
<td><strong>0.7192 (0.0154)</strong></td>
<td><strong>0.5480 (0.0440)</strong></td>
<td><strong>0.8583 (0.0083)</strong></td>
</tr>
</tbody>
</table>

*Missing samples*

## SSA of Daily Features

<table>
<thead>
<tr>
<th>Feature Extraction</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without SSA Denoising</td>
<td>0.6831 (0.0086)</td>
<td>0.5923 (0.0224)</td>
<td>0.3720 (0.0402)</td>
<td>0.8583 (0.0083)</td>
</tr>
<tr>
<td>With SSA</td>
<td><strong>0.7326 (0.0074)</strong></td>
<td><strong>0.7192 (0.0154)</strong></td>
<td><strong>0.5480 (0.0440)</strong></td>
<td><strong>0.8583 (0.0083)</strong></td>
</tr>
</tbody>
</table>

*Noisy, missing daily features*
Robust Prediction of Surgical Complications

- Machine learning models outperform standard surgical risk scores.
  - x2 AUPRC
  - x3 sensitivity at the same specificity

<table>
<thead>
<tr>
<th>Data Source</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random weighted classifier</td>
<td>0.5097 (0.0585)</td>
<td>0.4322 (0.0469)</td>
<td>0.1520 (0.0854)</td>
<td>0.8583 (0.0504)</td>
</tr>
<tr>
<td>NSQIP with Clinical Characteristics</td>
<td>0.6114 (0.0000)</td>
<td>0.4075 (0.0000)</td>
<td>0.2800 (0.0000)</td>
<td>0.8571 (0.0000)</td>
</tr>
<tr>
<td>ML with Clinical Characteristics</td>
<td>0.7632 (0.0085)</td>
<td>0.7374 (0.0206)</td>
<td>0.5800 (0.0699)</td>
<td>0.8583 (0.0083)</td>
</tr>
<tr>
<td>Wearable Data</td>
<td>0.7326 (0.0074)</td>
<td>0.7192 (0.0154)</td>
<td>0.5480 (0.0440)</td>
<td>0.8583 (0.0083)</td>
</tr>
<tr>
<td>Clinical Characteristics + Wearable Data</td>
<td>0.8802 (0.0050)</td>
<td>0.8871 (0.0087)</td>
<td>0.8320 (0.0160)</td>
<td>0.8583 (0.0083)</td>
</tr>
</tbody>
</table>

- NSQIP: American College of Surgeons National Surgical Quality Improvement Program
- AUROC: Area Under the Receiver Operating Characteristic Curve
- AUPRC: Area Under the Precision-Recall Curve
Robust Prediction of Surgical Complications

- Machine learning models outperform traditional surgical risk scores.
- Wearables & clinical characteristics provide complementary information.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random weighted classifier</td>
<td>0.5097 (0.0585)</td>
<td>0.4322 (0.0469)</td>
<td>0.1520 (0.0854)</td>
<td>0.8583 (0.0504)</td>
</tr>
<tr>
<td>NSQIP with Clinical Characteristics</td>
<td>0.6114 (0.0000)</td>
<td>0.4075 (0.0000)</td>
<td>0.2800 (0.0000)</td>
<td>0.8571 (0.0000)</td>
</tr>
<tr>
<td>ML with Clinical Characteristics</td>
<td>0.7632 (0.0085)</td>
<td>0.7374 (0.0206)</td>
<td>0.5800 (0.0699)</td>
<td>0.8583 (0.0083)</td>
</tr>
<tr>
<td>Wearable Data</td>
<td>0.7326 (0.0074)</td>
<td>0.7192 (0.0154)</td>
<td>0.5480 (0.0440)</td>
<td>0.8583 (0.0083)</td>
</tr>
<tr>
<td>Clinical Characteristics + Wearable Data</td>
<td><strong>0.8802 (0.0050)</strong></td>
<td><strong>0.8871 (0.0087)</strong></td>
<td><strong>0.8320 (0.0160)</strong></td>
<td><strong>0.8583 (0.0083)</strong></td>
</tr>
</tbody>
</table>

- NSQIP: American College of Surgeons National Surgical Quality Improvement Program
- AUROC: Area Under the Receiver Operating Characteristic Curve
- AUPRC: Area Under the Precision-Recall Curve
Extract **high-level** features from **fine-grained** wearable data

- Raw time series (minute) \(\rightarrow\) Daily features
- Time series of daily features \(\rightarrow\) High-level features

Extract **robust** features from **noisy** and **incomplete** wearable data

- Standardized daily features from incomplete time series
- Singular spectrum analysis for high-level feature extraction

---

Predict Clinical Outcomes with Wearables

- Complications after pancreatic surgery
  - Robust machine learning based on a small patient cohort

- Mental health disorders in the community
  - Deep learning for a large and diverse cohort

- Treatment response to depression therapy
  - Personalized prediction of treatment response
Mental Health Crisis

- Mental health disorders are prevalent.
  - ~3.8% of the population (280 million) experience depression (WHO).

- Over 50% of patients are not recognized or treated.

- Clinical visit is time-consuming and expensive.
  - Hindering timely diagnosis and intervention

- Detect mental health disorders with wearables devices?
  - Unobtrusive, multi-modal sensing
  - Activities, heart rate, and sleep are associated with mental health
Need: Large and Diverse Cohorts

- Mental health studies with wearables: small cohorts with limited diversity
  - 18 pregnant women [UbiComp’19]
  - 48-652 college students [UbiComp’14, UbiComp’19, J. Biomed. Inform. ‘22]
  - 1,002 healthy subjects [NPJ Digit. Med. ‘18]

- Machine learning approaches
  - Shallow models rely on ad hoc feature engineering.
  - Deep models tend to overfit.

- Small datasets limit the rigor and generality of results!
Deep Learning on a Large and Diverse Dataset

- **8,996** participants of the NIH All of Us program
  - **1,247** with depressive/anxiety disorders
  - **Wearable data** and **mental health diagnosis**
  - **Diverse population**

- **WearNet**: deep model for detecting mental health disorders with wearables

Joint work with Thomas Kannampallil (Informatics), Laura Jean Bierut (Psychiatry), Ruixuan Dai (CSE)
Outcomes and Inputs

- Mental health disorders
  - 20 depression and anxiety disorder diagnoses identified by clinicians
  - Positive label: a participant with any of the diagnoses

- Input features
  - Wearable data in a 60-day window
    - 10 daily features of step and heart rate
  - Static data from electrical health records and surveys
    - Age, race, ethnicity, gender, education, smoke history, alcohol history
WearNet

- Combining transformer encoder and convolutional neural network

- Transformer encoder (multi-head self-attention)
  - Identifies patterns across multiple timestamps

- Convolutional neural network (1-d)
  - Integrates neighborhood patterns

- Global max pooling
  - Identifies one global pattern for robustness

- Integrate wearable and static data at the top
  - Captures the underlying characteristics
### Detection Performance

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>AUROC</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow Models</td>
<td>LR</td>
<td>0.701(0.000)</td>
<td>0.351(0.000)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.592(0.000)</td>
<td>0.290(0.000)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.661(0.005)</td>
<td>0.349(0.007)</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.685(0.001)</td>
<td>0.365(0.000)</td>
</tr>
<tr>
<td>Deep Models</td>
<td>bi-LSTM</td>
<td>0.702(0.015)</td>
<td>0.464(0.011)</td>
</tr>
<tr>
<td></td>
<td>BRITS</td>
<td>0.693(0.012)</td>
<td>0.445(0.011)</td>
</tr>
<tr>
<td></td>
<td>CrossNet</td>
<td>0.682(0.021)</td>
<td>0.429(0.014)</td>
</tr>
<tr>
<td></td>
<td>TCN</td>
<td>0.629(0.021)</td>
<td>0.235(0.024)</td>
</tr>
<tr>
<td></td>
<td>Informer</td>
<td>0.705(0.008)</td>
<td>0.428(0.011)</td>
</tr>
<tr>
<td></td>
<td>WearNet</td>
<td>0.717(0.009)</td>
<td>0.487(0.008)</td>
</tr>
</tbody>
</table>

- **Shallow** models with feature engineering are less predictive.
- **RNN/TCN** models underperform **Transformer** models.
- **WearNet** achieves the best predictive performance.
Assign importance scores to features by approximating the integral of gradients.

- **Total steps** is the most important wearable feature.
- **Women** and **frequent smokers** are more likely to be diagnosed.
Mental health disorders are prevalent but significantly underdiagnosed.

A large and diverse cohort with 8,996 participants from All of Us

WearNet: deep model for detecting mental health disorders with wearables

An unobtrusive pathway for screening mental health disorders

R. Dai, T Kannampallil, S. Kim, V. Thornton, L. Bierut, C. Lu, Detecting Mental Disorders with Wearables: A Large Cohort Study, ACM/IEEE Conference on Internet of Things Design and Implementation (IoTDI'23), May 2023. Best Paper Award for IoT Data Analytics
Predict Clinical Outcomes with Wearables

- Complications after pancreatic surgery
  - Robust machine learning based on a small patient cohort

- Mental health disorders in the community
  - Deep learning for a large and diverse cohort

- Treatment response to depression therapy
  - Personalized prediction of treatment response
Personalized Prediction of Treatment Response

- Statistical analysis → population-level effectiveness of treatment
- Personalized prediction → precision medicine
- Machine learning from RCT data
  - Clinical (baseline): age, anxiety…
  - Fitbit (2 months): heart rate, sleep
  - Depression outcome (at 6 month)

Randomized Controlled Trial of Depression Therapy

106 patients

Group randomization

6-month trial period

- Intervention
  - Behavior therapy
  - Positive outcome
  - Negative outcome

- Control
  - No treatment

Baseline clinical measurements

Continuous wearable data

Joint work with Thomas Kannampallil (Informatics), Jun Ma (Medicine, UIC), Ruixuan Dai, Jingwen Zhang (CSE)
Predicting Personalized Treatment Response

➢ **Group-specific models**
  - Intervention group
  - Control group
  - Treatment-specific model
  - Risk model
  - Outcome w/ therapy
  - Outcome w/o therapy

- **Small groups → overfitting**

➢ **Multi-task learning (MTL) with a unified model**
  - MTL model
  - Outcome w/ therapy
  - Outcome w/o therapy

- **Exploit similarity**
- **Capture differences**
Multi-Task Learning Architecture

Learn commonality (data before treatment)

Clinical data are statistically similar

Common layer

Clinical data @ baseline

Wearable Data (Intervention)

Capture differences (data under treatment)

Wearable Data (Control)

Depression Intervention group

Depression Control group

Group-specific layer

Wearable data of the intervention group reflect effects of therapy

IoMT for Precision Medicine

- **Reliable** prediction of clinical outcomes
  - Wearable data are **fine-grained, noisy, incomplete**

- **Scaling** predictive models
  - Small cohorts: **robust** models with limited data
  - Large population: exploit **deep** learning models

- Prediction of **treatment response**
  - Support personalized **intervention**
Embed AI and Wearables in Healthcare

- Significant **clinical information** can be learned from wearables.
- Precision medicine: **personalized** prediction → **targeted** intervention
- Rigorous **clinical studies** are needed to validate AI models.
- Close **collaboration** between AI and clinical researchers is essential.
Multidisciplinary Team

- AI: Ruixuan Dai, Dingwen Li, Jingwen Zhang
- Health: Chet Hammill (Surgery), Thomas Kannampallil (Informatics), Laura Jean Bierut (Psychiatry), Jun Ma (Medicine)

The AI for Health Institute will bring together AI researchers and clinical investigators to forge new paths to solve significant health problems with advanced data-driven tools.

https://aihealth.wustl.edu/