AI for Health with Wearables

Chenyang Lu
AI for Health Institute (AIHealth)
Wearables

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

Unprecedented monitoring capability outside hospitals!
IoMT: Internet of Medical Things

- **Wearables**: wristband, watch, ring...
  - Long-term, non-obtrusive monitoring

- **Connectivity**: Bluetooth, WiFi, cellular
  - Real-time monitoring and intervention

- **Cloud**: computing and storage
  - Scalable to large population

- **Analytics**: machine learning
  - Predict outcomes and support intervention
IoMT for Precision Medicine

- **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

---

**Data Collection**

- Smart Phone
- Fitbit
- Fitbit Cloud
- Poll request
- Transfer data

**Machine Learning**

- Imputation for short missing segments
- Daily feature extraction
- High-level feature extraction
- Daily number of daily features
- N: # of days
- D: # of daily features
- Clinical features
Wearable Data

- Wristband provides time series data: step count, heart rate, and sleep stage.

**Fine-grained, noisy, incomplete time-series**
Wearable Data $\rightarrow$ 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**

![Example of step and heart rate data collected by Fitbit](image)
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**: pre-habilitation

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 complex (deep) models
- Extract features from wearable time series + shallow machine learning models
- Integrate wearable data and patient clinical characteristics
Robust Feature Engineering Pipeline

- **Daily Feature Extraction**
  - \( D \): Number of Daily Features
  - \( N \): Number of Days

- **High-level Feature Extraction**

**Inputs to predictive model**

- **Static Clinical Features**

**Wearable Time Series**

- Step Count
- Heart Rate
- Sleep Stage
Raw Time Series $\rightarrow$ Daily Features

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

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

- **Standardization**: account for missing data

Diagram:

1. Daily sedentary time
2. Duration of wearing the device
3. Daily sedentary time

#sample
Daily Features → High-level Features

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

Trend Extraction with Complete Data (30 data samples)
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.
  - impute missing daily features.

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)](image)

- mean
- standard deviation
- slope
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.
- **Wearable data + clinical characteristics → best predictive performance**

<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
## 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>0.7326 (0.0074)</td>
<td>0.7192 (0.0154)</td>
<td>0.5480 (0.0440)</td>
<td>0.8583 (0.0083)</td>
</tr>
</tbody>
</table>

### 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>0.7326 (0.0074)</td>
<td>0.7192 (0.0154)</td>
<td>0.5480 (0.0440)</td>
<td>0.8583 (0.0083)</td>
</tr>
</tbody>
</table>
Summary: Predict Postoperative Complications

- 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
- Including **1,247** with depressive/anxiety disorders
- Longitudinal wearable data and mental health diagnosis

- **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 depressive/anxiety disorder diagnoses identified by clinical experts
- 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
    - e.g., total steps, average heart rate, sedentary minutes
- Static data from electrical health records and surveys
  - Age, race, ethnicity, gender, education, smoke history, alcohol history
WearNet: Deep Model for Detecting Mental Health Disorders

- 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>

- **AUROC**: Area Under the Receiver Operating Characteristic Curve
- **AUPRC**: Area Under the Precision-Recall Curve

- **Shallow models** with feature engineering are less predictive.
- **RNN and TCN models** underperform Transformer models.
- **WearNet** achieves the best predictive performance.
Model Explanation

- 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.
Summary: Detect Mental Health Disorders

- 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.

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

- **Group randomization**
  - Random split (71:35)
  - 106 patients

- **6-month trial period**
  - **Intervention**
  - Behavior therapy
  - **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

Multi-task learning (MTL) with a unified model

- MTL model

- Outcome w/ therapy
- Outcome w/o therapy

- Small groups → overfitting
- Exploit similarity
- Capture differences
Multi-Task Learning Architecture

- **Clinical data** at baseline
  - Statistically similar between groups at baseline

- **Wearable data** during the first 2 months
  - Data of the intervention group reflect effects of therapy

Learn commonality *(data before treatment)*

Wearable Data (Intervention)

<table>
<thead>
<tr>
<th>Group</th>
<th>Clinical data @ baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention group</td>
<td></td>
</tr>
<tr>
<td>Control group</td>
<td></td>
</tr>
</tbody>
</table>

Capture differences *(data under treatment)*

- Depression
  - Intervention group
  - Control group
Summary: Predict Treatment Response

- **Personalized** predictions of treatment response → precision medicine
- Learn from data collected in **randomized controlled trials**
- Multi-task learning exploits the **commonalities** while capturing the **differences** between intervention and control groups

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 IoMT in Healthcare

- Significant **clinical information** can be learned from wearables.
- **AI** is key for extracting clinical information from wearable data.
- Rigorous **clinical studies** are needed to validate IoMT models.
- Precision medicine: **personalized prediction → intervention → outcome**
- Close **collaboration** between AI and clinical researchers is essential.
Interdisciplinary Team

- PhD students: Ruixuan Dai, Dingwen Li, Jingwen Zhang

- Clinical collaborators: Chet Hammill (Surgery), Thomas Kannampallil (Informatics), Laura Jean Bierut (Psychiatry), Jun Ma (Medicine)

- AI for Health Institute (AIHealth)

AIHealth: https://aihealth.wustl.edu/
Research: https://www.cse.wustl.edu/~lu/iomt.html