FIDDLE - Flexible Data-Driven pipeline for EHR preprocessing

**Motivation & Background**

**EHR Data**

**Challenges**
- Messy: high-dimensional, irregularly-sampled, multiple data types with frequent missingness
- Many decisions involved, labor-intensive, error-prone, and ad-hoc
- Heterogeneity makes comparisons difficult

**Goals:**
- Speed up and standardize EHR data preprocessing
- An open-source, generalizable, data-driven pipeline
- Offer a quick and reasonable starting point to build upon

**Dataset & Tasks**

**Widely shared EHR database:** MIMIC-III

**Task**

- In-hospital mortality: using the first 48hs of a patient's ICU visit (Harutyunyan et al., 2017)
- Acute Respiratory Failure (ARF): need for respiratory support with positive pressure mechanical ventilation (Stefan et al., 2013)
- Shock: inadequate perfusion; receipt of vasopressor therapy (Avni et al., 2015)

<table>
<thead>
<tr>
<th>Task</th>
<th>In-hospital mortality (48h)</th>
<th>ARF (4h)</th>
<th>ARF (12h)</th>
<th>shock (4h)</th>
<th>shock (12h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>8,577</td>
<td>15,873</td>
<td>14,174</td>
<td>19,342</td>
<td>17,588</td>
</tr>
<tr>
<td>Sensitive</td>
<td>12.0</td>
<td>18.3</td>
<td>9.7</td>
<td>14.9</td>
<td>7.8</td>
</tr>
</tbody>
</table>

**Patient risk stratification as binary classification**

Increasing Risk

**Three adverse outcomes, five prediction tasks**
- In-hospital mortality: using the first 48hs of a patient's ICU visit (Harutyunyan et al., 2017)
- Acute Respiratory Failure (ARF): need for respiratory support with positive pressure mechanical ventilation (Stefan et al., 2013)
- Shock: inadequate perfusion; receipt of vasopressor therapy (Avni et al., 2015)

**FIDDLE - Flexible Data-Driven pipeline for EHR preprocessing**

**Experimental Results**

**Results: Extracted Features**
- Extracted 4,414-7,508 features on 8,577-17,588 ICU stays in 30-150 minutes

**Time-invariant**
- Logistic Regression
- Random Forest
- 1D CNN
- LSTM

**Time-dependent**
- 1.5 hours: mean age
- 2.5 hours: daily insulin
- 5.5 hours: calculator

**AUCROCs (95% confidence intervals)**

<table>
<thead>
<tr>
<th>Method</th>
<th>In-hospital mortality (48h)</th>
<th>ARF (4h)</th>
<th>ARF (12h)</th>
<th>Shock (4h)</th>
<th>Shock (12h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.859</td>
<td>0.630</td>
<td>0.628</td>
<td>0.677</td>
<td>0.682</td>
</tr>
<tr>
<td>(0.799, 0.977)</td>
<td>(0.634, 0.667)</td>
<td>(0.588, 0.666)</td>
<td>(0.644, 0.711)</td>
<td>(0.740, 0.805)</td>
<td>(0.648, 0.721)</td>
</tr>
<tr>
<td>FIDDLE-LR</td>
<td>0.856</td>
<td>0.733</td>
<td>0.755</td>
<td>0.779</td>
<td>0.793</td>
</tr>
<tr>
<td>(0.821, 0.898)</td>
<td>(0.709, 0.773)</td>
<td>(0.717, 0.795)</td>
<td>(0.740, 0.865)</td>
<td>(0.715, 0.825)</td>
<td>(0.703, 0.887)</td>
</tr>
<tr>
<td>FIDDLE-RF</td>
<td>0.834</td>
<td>0.733</td>
<td>0.755</td>
<td>0.773</td>
<td>0.793</td>
</tr>
<tr>
<td>(0.790, 0.847)</td>
<td>(0.707, 0.773)</td>
<td>(0.723, 0.793)</td>
<td>(0.740, 0.865)</td>
<td>(0.715, 0.825)</td>
<td>(0.703, 0.887)</td>
</tr>
<tr>
<td>FIDDLE-CNN</td>
<td>0.886</td>
<td>0.751</td>
<td>0.786</td>
<td>0.794</td>
<td>0.794</td>
</tr>
<tr>
<td>(0.854, 0.916)</td>
<td>(0.718, 0.793)</td>
<td>(0.732, 0.805)</td>
<td>(0.747, 0.860)</td>
<td>(0.747, 0.860)</td>
<td>(0.747, 0.860)</td>
</tr>
<tr>
<td>FIDDLE-LSTM</td>
<td>0.886</td>
<td>0.744</td>
<td>0.767</td>
<td>0.777</td>
<td>0.794</td>
</tr>
<tr>
<td>(0.850, 0.915)</td>
<td>(0.720, 0.777)</td>
<td>(0.732, 0.800)</td>
<td>(0.747, 0.860)</td>
<td>(0.747, 0.860)</td>
<td>(0.747, 0.860)</td>
</tr>
</tbody>
</table>

**FIDDLE - an open-source pipeline for EHR preprocessing**

tiny.cc/get_FIDDLE