End-to-End Machine Learning Frameworks for Medicine: Imputation, Interpretation and Synthetic Data Generation

Jinsung Yoon

Advisor: Prof. Mihaela van der Schaar

April 13th, 2020
Tremendous Success of Machine Learning

Convolutional Neural Networks for Object Detection

Generative Adversarial Networks for Data Generation

Transformer for Language Translation
Rapid Increase of Electronic Health Records

Intensive Care Unit Database (30k)

Primary Care Hospital Database (500k)

Genomic Database (8000k)

COVID-19 Database
Objective: Machine Learning for Medicine

Electronic Health Records

Machine Learning Models

Clinical Decision Supports
Challenges: (1) Missing Data

Different measurements across hospitals

Irregular sampling

Doctor bias
Challenges: (2) Model Interpretation

Model

Data and Prediction

Explanation

Human makes decision

<table>
<thead>
<tr>
<th>Interpretation Scope</th>
<th>Global</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td>Decision Tree</td>
<td>Attention Mechanism</td>
</tr>
<tr>
<td>(a)</td>
<td>(b)</td>
<td></td>
</tr>
<tr>
<td>Intrisic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posthoc</td>
<td>Mimic Learning</td>
<td>Instance Heatmap</td>
</tr>
<tr>
<td>(c)</td>
<td>(d)</td>
<td></td>
</tr>
</tbody>
</table>
Challenges: (3) Privacy

Impossible to share directly

Hospitals → Private medical data → Strong regularization

HIPAA / GDPR

ML community
Part I: Machine Learning Pipeline for Healthcare

Challenge 1: Missing Data

Data Imputation

Model Training

Prediction

Challenge 2: Interpretation

Model Interpretation
Part II: Synthetic Data Generation

• **Enable to share** the private medical data (by sharing de-identified synthetic data) to **machine learning community** for developing machine learning tools easier.
Outline

• Part I: Machine Learning Pipeline for Healthcare
  • Data Imputation
    • Section 1: Generative Adversarial Imputation Network (GAIN)
    • Section 2: Multi-directional RNN (MRNN)
  • Model Interpretation
    • Section 3: Instance-wise Variable Selection (INVASE)
  • Real-world Application
    • Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)

• Part II: Synthetic Data Generation
  • Section 5: Time-series Generative Adversarial Nets (TimeGAN)
Outline

• Part I: Machine Learning Pipeline for Healthcare
  • Data Imputation
    • Section 1: Generative Adversarial Imputation Network (GAIN)
    • Section 2: Multi-directional RNN (MRNN)
  • Model Interpretation
    • Section 3: Instance-wise Variable Selection (INVASE)
  • Real-world Application
    • Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)
  • Part II: Synthetic Data Generation
    • Section 5: Time-series Generative Adversarial Nets (TimeGAN)
Section 1: Data Imputation

Missing data is a ubiquitous problem:
- Most medical datasets contain **many missing values**.
- Most machine learning models **cannot utilize datasets** with missing values.
- **Handling missing values is critical** for accurate diagnosis, prognosis and treatment.
Section 1: Data Imputation

**Solution:**
*Impute* the missing data to *complete* the dataset.

**What is good imputation?**
*Imputed values* follow the *same underlying data distribution*.
Real-world Medical Example

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>64</td>
<td>Yes</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Patient 2</td>
<td>86</td>
<td>?</td>
<td>143</td>
<td>?</td>
</tr>
<tr>
<td>Patient 3</td>
<td>52</td>
<td>No</td>
<td>?</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>64</td>
<td>Yes</td>
<td>131</td>
<td>No</td>
</tr>
<tr>
<td>Patient 2</td>
<td>86</td>
<td>Yes</td>
<td>143</td>
<td>Yes</td>
</tr>
<tr>
<td>Patient 3</td>
<td>52</td>
<td>No</td>
<td>115</td>
<td>No</td>
</tr>
</tbody>
</table>
Intuition: Good vs Bad Imputation

We model this **intuition** using **Generative Adversarial Networks**. We called our model **Generative Adversarial Imputation Networks (GAIN)**.

### Good Example

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient 1</strong></td>
<td>64</td>
<td>Yes</td>
<td>131</td>
<td>No</td>
<td>29</td>
</tr>
<tr>
<td><strong>Patient 2</strong></td>
<td>86</td>
<td>Yes</td>
<td>143</td>
<td>Yes</td>
<td>33</td>
</tr>
<tr>
<td><strong>Patient 3</strong></td>
<td>52</td>
<td>No</td>
<td>115</td>
<td>No</td>
<td>25</td>
</tr>
</tbody>
</table>

### Bad Example

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient 1</strong></td>
<td>64</td>
<td>Yes</td>
<td>237</td>
<td>No</td>
<td>29</td>
</tr>
<tr>
<td><strong>Patient 2</strong></td>
<td>86</td>
<td>Yes</td>
<td>143</td>
<td>Yes</td>
<td>33</td>
</tr>
<tr>
<td><strong>Patient 3</strong></td>
<td>52</td>
<td>No</td>
<td>15</td>
<td>No</td>
<td>100</td>
</tr>
</tbody>
</table>

Background: Generative Adversarial Networks

- Generative models to estimate the joint distribution of features.
- Consists of generator and discriminator that are trained in an adversarial way.

\[
\min_G \max_D V(D, G)
\]

\[
V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]
\]
Components of Data Matrix

### Given Data Matrix

<table>
<thead>
<tr>
<th></th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
<td>?</td>
<td>?</td>
<td>( x_{15} )</td>
</tr>
<tr>
<td>#2</td>
<td>( x_{21} )</td>
<td>?</td>
<td>( x_{23} )</td>
<td>?</td>
<td>( x_{25} )</td>
</tr>
<tr>
<td>#3</td>
<td>( x_{31} )</td>
<td>( x_{32} )</td>
<td>?</td>
<td>( x_{34} )</td>
<td>?</td>
</tr>
</tbody>
</table>

### Observed Data Matrix

<table>
<thead>
<tr>
<th></th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
<td>0</td>
<td>0</td>
<td>( x_{15} )</td>
</tr>
<tr>
<td>#2</td>
<td>( x_{21} )</td>
<td>0</td>
<td>( x_{23} )</td>
<td>0</td>
<td>( x_{25} )</td>
</tr>
<tr>
<td>#3</td>
<td>( x_{31} )</td>
<td>( x_{32} )</td>
<td>0</td>
<td>( x_{34} )</td>
<td>0</td>
</tr>
</tbody>
</table>

### Mask Matrix (\( M \))

<table>
<thead>
<tr>
<th></th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Generator of GAIN

**Observed Data Matrix**

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>0</td>
<td>0</td>
<td>$x_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$x_{21}$</td>
<td>0</td>
<td>$x_{23}$</td>
<td>0</td>
<td>$x_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>0</td>
<td>$x_{34}$</td>
<td>0</td>
</tr>
</tbody>
</table>

**Mask Matrix ($M$)**

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Noise Matrix ($Z$)**

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>0</td>
<td>0</td>
<td>$z_{13}$</td>
<td>$z_{14}$</td>
<td>0</td>
</tr>
<tr>
<td>#2</td>
<td>0</td>
<td>$z_{22}$</td>
<td>0</td>
<td>$z_{24}$</td>
<td>0</td>
</tr>
<tr>
<td>#3</td>
<td>0</td>
<td>0</td>
<td>$z_{33}$</td>
<td>0</td>
<td>$z_{35}$</td>
</tr>
</tbody>
</table>

**Generated Data Matrix ($\tilde{x}$)**

<table>
<thead>
<tr>
<th></th>
<th>$\tilde{x}_{11}$</th>
<th>$\tilde{x}_{12}$</th>
<th>$\tilde{x}_{13}$</th>
<th>$\tilde{x}_{14}$</th>
<th>$\tilde{x}_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$\tilde{x}_{21}$</td>
<td>$\tilde{x}_{22}$</td>
<td>$\tilde{x}_{23}$</td>
<td>$\tilde{x}_{24}$</td>
<td>$\tilde{x}_{25}$</td>
</tr>
<tr>
<td>#2</td>
<td>$\tilde{x}_{31}$</td>
<td>$\tilde{x}_{32}$</td>
<td>$\tilde{x}_{33}$</td>
<td>$\tilde{x}_{34}$</td>
<td>$\tilde{x}_{35}$</td>
</tr>
</tbody>
</table>

Generator (G)
**Generator of GAIN**

### Generated Data Matrix ($\tilde{X}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$\tilde{x}_{11}$</td>
<td>$\tilde{x}_{12}$</td>
<td>$\tilde{x}_{13}$</td>
<td>$\tilde{x}_{14}$</td>
<td>$\tilde{x}_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$\tilde{x}_{21}$</td>
<td>$\tilde{x}_{22}$</td>
<td>$\tilde{x}_{23}$</td>
<td>$\tilde{x}_{24}$</td>
<td>$\tilde{x}_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$\tilde{x}_{31}$</td>
<td>$\tilde{x}_{32}$</td>
<td>$\tilde{x}_{33}$</td>
<td>$\tilde{x}_{34}$</td>
<td>$\tilde{x}_{35}$</td>
</tr>
</tbody>
</table>

### Imputed Data Matrix ($\hat{X}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$\tilde{x}_{13}$</td>
<td>$\tilde{x}_{14}$</td>
<td>$x_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$x_{21}$</td>
<td>$\tilde{x}_{22}$</td>
<td>$x_{23}$</td>
<td>$\tilde{x}_{24}$</td>
<td>$x_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$\tilde{x}_{33}$</td>
<td>$x_{34}$</td>
<td>$\tilde{x}_{35}$</td>
</tr>
</tbody>
</table>
Discriminator of GAIN

Imputed Data Matrix ($\tilde{X}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$\tilde{x}_{13}$</td>
<td>$\tilde{x}_{14}$</td>
<td>$x_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$x_{21}$</td>
<td>$\tilde{x}_{22}$</td>
<td>$x_{23}$</td>
<td>$\tilde{x}_{24}$</td>
<td>$x_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$\tilde{x}_{33}$</td>
<td>$x_{34}$</td>
<td>$\tilde{x}_{35}$</td>
</tr>
</tbody>
</table>

Discriminator (D)

Estimated Mask Matrix ($\hat{M}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$\hat{m}_{11}$</td>
<td>$\hat{m}_{12}$</td>
<td>$\hat{m}_{13}$</td>
<td>$\hat{m}_{14}$</td>
<td>$\hat{m}_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$\hat{m}_{21}$</td>
<td>$\hat{m}_{22}$</td>
<td>$\hat{m}_{23}$</td>
<td>$\hat{m}_{24}$</td>
<td>$\hat{m}_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$\hat{m}_{31}$</td>
<td>$\hat{m}_{32}$</td>
<td>$\hat{m}_{33}$</td>
<td>$\hat{m}_{34}$</td>
<td>$\hat{m}_{35}$</td>
</tr>
</tbody>
</table>

Ground Truth

Mask Matrix ($M$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
GAIN vs GAN

**GAIN**

\[
\begin{array}{ccccc}
f_1 & f_2 & f_3 & f_4 & f_5 \\
\hline
\#1 & x_{11} & x_{12} & \tilde{x}_{13} & \tilde{x}_{14} & x_{15} \\
\#2 & x_{21} & \tilde{x}_{22} & x_{23} & \tilde{x}_{24} & x_{25} \\
\#3 & x_{31} & x_{32} & \tilde{x}_{33} & x_{34} & \tilde{x}_{35} \\
\end{array}
\]

**Discriminator (D)**

\[
\begin{array}{ccccc}
\hat{m}_1 & \hat{m}_2 & \hat{m}_3 & \hat{m}_4 & \hat{m}_5 \\
\#1 & \hat{m}_{11} & \hat{m}_{12} & \hat{m}_{13} & \hat{m}_{14} & \hat{m}_{15} \\
\#2 & \hat{m}_{21} & \hat{m}_{22} & \hat{m}_{23} & \hat{m}_{24} & \hat{m}_{25} \\
\#3 & \hat{m}_{31} & \hat{m}_{32} & \hat{m}_{33} & \hat{m}_{34} & \hat{m}_{35} \\
\end{array}
\]

**Standard GAN**

\[
\begin{array}{ccccc}
f_1 & f_2 & f_3 & f_4 & f_5 \\
\hline
\#1 & x_{11} & x_{12} & \tilde{x}_{13} & \tilde{x}_{14} & x_{15} \\
\#2 & x_{21} & \tilde{x}_{22} & x_{23} & \tilde{x}_{24} & x_{25} \\
\#3 & x_{31} & x_{32} & \tilde{x}_{33} & x_{34} & \tilde{x}_{35} \\
\end{array}
\]

**Discriminator (D)**

\[
\begin{array}{ccccc}
\hat{p}_1 & \hat{p}_2 & \hat{p}_3 \\
\#1 & \hat{p}_1 \\
\#2 & \hat{p}_2 \\
\#3 & \hat{p}_3 \\
\end{array}
\]
Hint Mechanism

Imputed Data Matrix ($\hat{X}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$\bar{x}_{13}$</td>
<td>$\bar{x}_{14}$</td>
<td>$x_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$x_{21}$</td>
<td>$\bar{x}_{22}$</td>
<td>$x_{23}$</td>
<td>$\bar{x}_{24}$</td>
<td>$x_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$\bar{x}_{33}$</td>
<td>$x_{34}$</td>
<td>$\bar{x}_{35}$</td>
</tr>
</tbody>
</table>

Hint Matrix ($H$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Discriminator ($D$)

Estimated Mask Matrix ($\hat{M}$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>$\hat{m}_{11}$</td>
<td>$\hat{m}_{12}$</td>
<td>$\hat{m}_{13}$</td>
<td>$\hat{m}_{14}$</td>
<td>$\hat{m}_{15}$</td>
</tr>
<tr>
<td>#2</td>
<td>$\hat{m}_{21}$</td>
<td>$\hat{m}_{22}$</td>
<td>$\hat{m}_{23}$</td>
<td>$\hat{m}_{24}$</td>
<td>$\hat{m}_{25}$</td>
</tr>
<tr>
<td>#3</td>
<td>$\hat{m}_{31}$</td>
<td>$\hat{m}_{32}$</td>
<td>$\hat{m}_{33}$</td>
<td>$\hat{m}_{34}$</td>
<td>$\hat{m}_{35}$</td>
</tr>
</tbody>
</table>

Partial Information

Mask Matrix ($M$)

<table>
<thead>
<tr>
<th></th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
### Example on Hint Mechanism

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>Yes</td>
<td>100</td>
<td>Yes</td>
<td>33</td>
</tr>
<tr>
<td>83</td>
<td>No</td>
<td>143</td>
<td>No</td>
<td>23</td>
</tr>
</tbody>
</table>

#### Imputed Data Matrix ($\hat{X}$)

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Diabetes</td>
</tr>
<tr>
<td>64</td>
<td>Yes</td>
</tr>
<tr>
<td>83</td>
<td>No</td>
</tr>
</tbody>
</table>

#### Hint Matrix ($H$)

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetes</th>
<th>SBP</th>
<th>Hypertension</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Discriminator Decision

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Diabetes</td>
</tr>
<tr>
<td>64</td>
<td>Yes</td>
</tr>
<tr>
<td>83</td>
<td>No</td>
</tr>
</tbody>
</table>
GAIN: Entire Process

\[ V(D, G) = \mathbb{E}_{\hat{X}, M, H} \left[ M^T \log D(\hat{X}, H) + (1 - M)^T \log (1 - D(\hat{X}, H)) \right] \]

min max \( V(D, G) \)
Experiments: How to Evaluate Imputation Method?

If we have **complete data**, 
- Introduce missingness and evaluate **how well the imputation methods recover the original data**.

If we do **not have complete data**, 
- Evaluate **prediction accuracy after imputation**.
## Experiments: GAIN with Complete Data

**RMSE Imputation Performance**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Breast</th>
<th>Spam</th>
<th>Letter</th>
<th>Credit</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIN</td>
<td>.0546 ± .0006</td>
<td>.0513 ± .0016</td>
<td>.1198 ± .0005</td>
<td>.1858 ± .0010</td>
<td>.1441 ± .0007</td>
</tr>
<tr>
<td>MICE</td>
<td>.0646 ± .0028</td>
<td>.0699 ± .0010</td>
<td>.1537 ± .0006</td>
<td>.2585 ± .0011</td>
<td>.1763 ± .0007</td>
</tr>
<tr>
<td>MissForest</td>
<td>.0608 ± .0013</td>
<td>.0553 ± .0013</td>
<td>.1605 ± .0004</td>
<td>.1976 ± .0015</td>
<td>.1623 ± .012</td>
</tr>
<tr>
<td>Matrix</td>
<td>.0946 ± .0020</td>
<td>.0542 ± .0006</td>
<td>.1442 ± .0006</td>
<td>.2602 ± .0073</td>
<td>.2282 ± .0005</td>
</tr>
<tr>
<td>Auto-encoder</td>
<td>.0697 ± .0018</td>
<td>.0670 ± .0030</td>
<td>.1351 ± .0009</td>
<td>.2388 ± .0005</td>
<td>.1667 ± .0014</td>
</tr>
<tr>
<td>EM</td>
<td>.0634 ± .0021</td>
<td>.0712 ± .0012</td>
<td>.1563 ± .0012</td>
<td>.2604 ± .0015</td>
<td>.1912 ± .0011</td>
</tr>
</tbody>
</table>

GAIN **consistently outperforms** to state-of-the-art imputation methods across various datasets.
GAIN preserves the feature-label relationships better in terms of both AUROC and Congeniality.

*Congeniality: Compare the coefficients of logistic regressions using imputed data and original data
Part I: Machine Learning Pipeline for Healthcare

- **Data Imputation**
  - Section 1: Generative Adversarial Imputation Network (GAIN)
  - **Section 2**: Multi-directional RNN (MRNN)
- Model Interpretation
  - Section 3: Instance-wise Variable Selection (INVASE)
- Real-world Application
  - Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)

Part II: Synthetic Data Generation

- Section 5: Time-series Generative Adversarial Nets (TimeGAN)
Section 2: Data Imputation in Longitudinal Settings

- **Missing data** is also pervasive in **multivariate longitudinal** dataset.
- **Various sampling rates across** different streams (irregular sampling)
Learn the correlation in time and across features:

- **In time:** Interpolation
- **Across features:** Imputation
Sequentially learned the correlation in time and across features:

- **Interpolation** – using **Bi-RNN** to capture the correlation **in time**
- **Imputation** – using **Fully Connected Network** to capture the correlation **across the features**

J. Yoon, W. R. Zame and M. van der Schaar, "Estimating Missing Data in Temporal Data Streams Using Multi-directional Recurrent Neural Networks," IEEE Transactions on Biomedical Engineering, 2018
Background: Recurrent Neural Networks

Components:
- Input \( (x_t) \) – multi-variate time-series
- Hidden state \( (h_t) \) – Intermediate state, acting as “memory” of previous inputs
- Output \( (o_t) \) – time-series predictions.
Background: Bi-directional Recurrent Neural Networks

Additional components:

• Backward hidden layer ($\vec{h}_t$)

• Network can utilize both previous and future observations to predict the current time step

• Causal prediction is not necessary for imputation.
M-RNN Architecture: Overall
M-RNN Architecture: Bi-RNN

- Using modified Bi-RNN to capture the correlations in time.
- Note that the timing of the inputs is lagged in the forward direction and advanced in the backward direction.
M-RNN Architecture: Fully Connected Network

- Using **Fully Connected Network** to capture the correlations across the features.

- Using **dropout** for multiple imputations

Note that Bi-RNN and FCN are jointly optimized.
Experiments: How to Evaluate Imputation Method?

If we have **complete data**,  
- Introduce missingness and evaluate **how well the imputation methods recover the original data**.

If we do **not have complete data**,  
- Evaluate **prediction accuracy after imputation**.
### 5 Longitudinal Medical Datasets

**ICU data**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>MIMIC-III</th>
<th>Deterioration</th>
<th>UNOS-Heart</th>
<th>UNOS-Lung</th>
<th>Biobank</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Patients</td>
<td>23,160</td>
<td>6,094</td>
<td>69,205</td>
<td>32,986</td>
<td>3,902</td>
</tr>
<tr>
<td># of Dimensions (Cont, Cat)</td>
<td>40 (31, 9)</td>
<td>38 (16, 22)</td>
<td>34 (10, 24)</td>
<td>34 (10, 24)</td>
<td>113 (67, 46)</td>
</tr>
<tr>
<td>Label ((y = 1))</td>
<td>1,320 (5.7%)</td>
<td>306 (5.3%)</td>
<td>4,844 (7.0%)</td>
<td>2,276 (6.9%)</td>
<td>195 (5.0%)</td>
</tr>
<tr>
<td>Avg # of samples</td>
<td>24.3</td>
<td>34.3</td>
<td>6.2</td>
<td>4.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Avg missing rate</td>
<td>75.0%</td>
<td>61.4%</td>
<td>59.1%</td>
<td>58.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Avg measure Freq.</td>
<td>1 hr / 12 hrs</td>
<td>4 hrs / 24 hrs</td>
<td>1 year</td>
<td>1 year</td>
<td>2.3 years</td>
</tr>
<tr>
<td>Avg Corr within streams</td>
<td>0.4122</td>
<td>0.3436</td>
<td>0.1213</td>
<td>0.1157</td>
<td>0.2424</td>
</tr>
<tr>
<td>Avg Corr across streams</td>
<td>0.3127</td>
<td>0.3454</td>
<td>0.0875</td>
<td>0.0897</td>
<td>0.0506</td>
</tr>
</tbody>
</table>

**Ward data**

**Transplant data**

**Primary care hospital data**

**5 longitudinal medical datasets with different characteristics**
Experiments: M-RNN on Longitudinal Medical Data

**RMSE Imputation Performance**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MIMIC-III</th>
<th>Deterioration</th>
<th>UNOS-Heart</th>
<th>UNOS-Lung</th>
<th>Biobank</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-RNN (MI)</td>
<td>0.0141 (-)</td>
<td>0.0105 (-)</td>
<td>0.0479 (-)</td>
<td>0.0606 (-)</td>
<td>0.0637 (-)</td>
</tr>
<tr>
<td>M-RNN (SI)</td>
<td>0.0144 (-)</td>
<td>0.0108 (-)</td>
<td>0.0477 (-)</td>
<td>0.0609 (-)</td>
<td>0.0629 (-)</td>
</tr>
<tr>
<td>[CBS16]</td>
<td>0.0337 (58.2%)</td>
<td>0.0258 (59.3%)</td>
<td>0.1352 (64.6%)</td>
<td>0.1343 (54.9%)</td>
<td>0.0812 (21.6%)</td>
</tr>
<tr>
<td>[LKW16]</td>
<td>0.0295 (52.2%)</td>
<td>0.0241 (56.4%)</td>
<td>0.1179 (59.4%)</td>
<td>0.1264 (52.1%)</td>
<td>0.0801 (20.5%)</td>
</tr>
<tr>
<td>[CPC18]</td>
<td>0.0292 (51.7%)</td>
<td>0.0233 (54.9%)</td>
<td>0.1057 (54.7%)</td>
<td>0.1172 (48.3%)</td>
<td>0.0778 (18.1%)</td>
</tr>
<tr>
<td>Spline</td>
<td>0.0735 (80.8%)</td>
<td>0.0215 (51.2%)</td>
<td>0.1102 (56.5%)</td>
<td>0.1199 (49.5%)</td>
<td>0.0845 (24.6%)</td>
</tr>
<tr>
<td>Cubic</td>
<td>0.0279 (49.5%)</td>
<td>0.0223 (52.9%)</td>
<td>0.1072 (55.3%)</td>
<td>0.1177 (48.5%)</td>
<td>0.0887 (28.2%)</td>
</tr>
<tr>
<td>MICE</td>
<td>0.0611 (76.9%)</td>
<td>0.0319 (67.1%)</td>
<td>0.1147 (58.2%)</td>
<td>0.1151 (47.4%)</td>
<td>0.0915 (30.4%)</td>
</tr>
<tr>
<td>MissForest</td>
<td>0.0293 (51.9%)</td>
<td>0.0264 (60.2%)</td>
<td>0.0489 (2.0%)</td>
<td>0.0652 (7.1%)</td>
<td>0.0892 (28.6%)</td>
</tr>
<tr>
<td>EM</td>
<td>0.0467 (69.8%)</td>
<td>0.0355 (70.4%)</td>
<td>-</td>
<td>-</td>
<td>0.0978 (34.9%)</td>
</tr>
<tr>
<td>Matrix Completion</td>
<td>0.0311 (54.7%)</td>
<td>0.0264 (60.2%)</td>
<td>0.0974 (50.8%)</td>
<td>0.0942 (35.7%)</td>
<td>0.0886 (28.1%)</td>
</tr>
<tr>
<td>Auto-encoder</td>
<td>0.0412 (66.0%)</td>
<td>0.0309 (65.0%)</td>
<td>0.0589 (18.7%)</td>
<td>0.0712 (14.9%)</td>
<td>0.0805 (20.9%)</td>
</tr>
<tr>
<td>MCMC</td>
<td>0.0437 (67.7%)</td>
<td>0.0364 (71.2%)</td>
<td>0.1091 (56.1%)</td>
<td>0.1124 (46.1%)</td>
<td>0.0936 (31.9%)</td>
</tr>
</tbody>
</table>

**M-RNN consistently outperforms to state-of-the-art baselines across various medical datasets**
Experiments: Preserving Feature Label Relationship

Performance gain of better imputation method (M-RNN) increases as the missing rate increases.
Outline

• Part I: Machine Learning Pipeline for Healthcare
  • Data Imputation
    • Section 1: Generative Adversarial Imputation Network (GAIN)
    • Section 2: Multi-directional RNN (MRNN)
  • Model Interpretation
    • Section 3: Instance-wise Variable Selection (INVASE)
  • Real-world Application
    • Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)

• Part II: Synthetic Data Generation
  • Section 5: Time-series Generative Adversarial Nets (TimeGAN)
• What are the **reasons** that the trained model make a **certain prediction**?
Section 3: Model Interpretability

<table>
<thead>
<tr>
<th>Interpretation Manner</th>
<th>Global</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td><img src="image" alt="Decision Tree" /></td>
<td><img src="image" alt="Attention Mechanism" /></td>
</tr>
<tr>
<td>Posthoc</td>
<td><img src="image" alt="Mimic Learning" /></td>
<td><img src="image" alt="Instance Heatmap" /></td>
</tr>
</tbody>
</table>

- **Interpretation Scope**
  - **Global**
    - Root node
    - Split condition
  - **Local**
    - Input sequence
    - Output sequence
    - Weight

- **Decision Tree**
  - Leaf node (decision)

- **Attention Mechanism**
  - Heavy heated
  - Light heated
Instance-wise Feature Selection

• Understanding which features are relevant to an outcome for each sample is critical in improving predictions and interpretability.

• Propose a novel “instance-wise feature selection” which attempts to learn which subset of the features are relevant for each sample.

• We call this model as INVASE.

• **Different patients** have different feature importance.
• Therefore, doctors should focus on different measurements for different patients.
What INVASE can do?

- INVASE provides **individualized feature importance** for each prediction as the **explanations**.

- It can combine with **any black-box predictive models** to provide **interpretability**.
• **Selector** selects the subset of features among the entire features.
• **Predictor** predicts the label only with the selected features

• **Objective:** To **maximize** the prediction accuracy using **minimum** number of **selected features**
Problem Formulation

- **$x$:** Patient features
- **$Y$:** ‘Original patient outcomes’ or ‘black-box model outputs’
- **$S: \mathcal{X} \to \{0,1\}^d$:** Selector function,
- **$S(x)$:** Selected features

---

**Objective:** minimize $S(x)$

**Constraints:**

\[
(Y | X^{(S(x))} = x^{(S(x))}) = (Y | X = x)
\]

---

We would like to find the **selector function $S$** that minimizes the number of **selected features $S(x)$** that satisfies the **conditional distributions equality constraints**.
Using Lagrange multiplier, we convert the optimization problem as to minimize the summation of the number of selected features $S(x)$ and KL divergence between conditional distributions with and without feature selection.

$$\mathcal{L}(S) = \mathbb{E}[KL[(Y|X^{(S(x))} = x^{(S(x))})|| (Y|X = x)] + \lambda ||S(x)||]$$
INVASE Architecture

- Entire features are fed into the selector networks (actor) which outputs a vector of selection probabilities.
- The predictor network (critic) receives the selected features, makes prediction, and provide feedbacks to the actor.
INVASE: Predictor Network

- **Predictor:** $f^\phi : \mathcal{X} \times \{0, 1\}^d \to [0, 1]^c$

- **Loss Function for Predictor Network**

$$l_1(\phi) = -\mathbb{E}_{(x, y) \sim p, s \sim \pi_\theta(x, \cdot)} \left[ \sum_{i=1}^c y_i \log(f_i^\phi(x^{(s)}, s)) \right]$$
INVASE: Baseline Network

• **Baseline**: $f^\gamma: \mathcal{X} \rightarrow [0, 1]^c$

• **Loss Function for Baseline Network**

$$l_3(\gamma) = -\mathbb{E}_{(x, y) \sim p} \left[ \sum_{i=1}^c y_i \log(f_i^\gamma(x)) \right]$$

• **Overall Loss Function**:

$$\hat{l}(x, s) = -\left[ \sum_{i=1}^c y_i \log(f_i^\phi(x^{(s)}, s)) - \sum_{i=1}^c y_i \log(f_i^\gamma(x)) \right]$$
INVASE: Selector Network

- **Original Selector:** $S: \mathcal{X} \rightarrow \{0, 1\}^d$

- **Alternative Selector:** $\hat{S}^\theta: \mathcal{X} \rightarrow [0, 1]^d \; + \; \text{Sampler}$

- **Probability Mass Function of Sampler:**

  $$\pi_\theta(x, s) = \prod_{i=1}^d \hat{S}_i^\theta(x)^{s_i} (1 - \hat{S}_i^\theta(x))^{1-s_i}$$
INVASE: Selector Loss

\[ \hat{l}(x, s) = - \sum_{i=1}^{c} y_i \log(f_i^\phi(x^{(s)}, s)) - \sum_{i=1}^{c} y_i \log(f_i^\gamma(x)) \]

\[ l_2(\theta) = \mathbb{E}_{(x, y) \sim p} \left[ \mathbb{E}_{s \sim \pi_\theta(x, \cdot)} \left[ \hat{l}(x, s) + \lambda \|s\|_0 \right] \right] \]

\[ = \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \left( \sum_{s \in \{0, 1\}^d} \pi_\theta(x, s) \left( \hat{l}(x, s) + \lambda \|s\|_0 \right) \right) dx dy. \]
INVASE: Selector Loss

- Directly compute the gradient of the Selector Network Loss due to the non-differentiable sampling process.

\[
\nabla_\theta l_2(\theta) = \int_{X \times Y} \! p(x, y) \left( \sum_{s \in \{0,1\}^d} \nabla_\theta \pi_\theta(x, s) \left( \hat{l}(x, s) + \lambda ||s||_0 \right) \right) \, dx \, dy \\
= \int_{X \times Y} \! p(x, y) \left( \sum_{s \in \{0,1\}^d} \frac{\nabla_\theta \pi_\theta(x, s)}{\pi_\theta(x, s)} \pi_\theta(x, s) \left( \hat{l}(x, s) + \lambda ||s||_0 \right) \right) \, dx \, dy \\
= \int_{X \times Y} \! p(x, y) \left( \sum_{s \in \{0,1\}^d} \nabla_\theta \log \pi_\theta(x, s) \pi_\theta(x, s) \left( \hat{l}(x, s) + \lambda ||s||_0 \right) \right) \, dx \, dy \\
= \mathbb{E}_{(x,y) \sim p} \left[ \mathbb{E}_{s \sim \pi_\theta(x, \cdot)} \left[ \left( \hat{l}(x, s) + \lambda ||s||_0 \right) \nabla_\theta \log \pi_\theta(x, s) \right] \right].
\]
Experiments: How to Evaluate Interpretation Method?

If we have **ground truth feature importance**,  
- Evaluate whether the interpretation method can **discover the ground truth**.

If we do **not have ground truth feature importance**,  
- Evaluate **prediction accuracy** after instance-wise feature selection.

![Diagram of model training and prediction process](image_url)
With Ground Truth Feature Importance: Synthetic Data

- Syn1: \( \logit(X) = \exp(X_1 X_2) \)
- Syn2: \( \logit(X) = \exp(\sum_{i=3}^{6} X_i^2 - 4) \)
- Syn3: \( \logit(X) = -10 \sin 2X_7 + 2|X_8| + X_9 + \exp(-X_{10}) \)

- Syn4: If \( X_{11} < 0 \), logit follows Syn1, otherwise, logit follows Syn2
- Syn5: If \( X_{11} < 0 \), logit follows Syn1, otherwise, logit follows Syn3
- Syn6: If \( X_{11} < 0 \), logit follows Syn2, otherwise, logit follows Syn3

\[ X \sim \mathcal{N}(0, I) \]

\[ P(Y = 1 | X) = \frac{1}{1 + \logit(X)} \]
**Discovery on Synthetic Data**

**Important feature discovery** results for 6 Synthetic Datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Syn1</th>
<th>Syn2</th>
<th>Syn3</th>
<th>Syn4</th>
<th>Syn5</th>
<th>Syn6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics (%)</td>
<td>TPR</td>
<td>FDR</td>
<td>TPR</td>
<td>FDR</td>
<td>TPR</td>
<td>FDR</td>
</tr>
<tr>
<td>INVASE</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>92.0</td>
<td>0.0</td>
</tr>
<tr>
<td>L2X</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>69.4</td>
<td>30.6</td>
</tr>
<tr>
<td>LIME</td>
<td>13.8</td>
<td>86.2</td>
<td>100.0</td>
<td>0.0</td>
<td>98.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Shapley</td>
<td>60.4</td>
<td>39.6</td>
<td>93.3</td>
<td>6.7</td>
<td>90.9</td>
<td>9.1</td>
</tr>
<tr>
<td>Knockoff</td>
<td>10.0</td>
<td>70.0</td>
<td>8.7</td>
<td>36.2</td>
<td>81.2</td>
<td>17.5</td>
</tr>
<tr>
<td>Tree</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>SCFS</td>
<td>23.5</td>
<td>76.5</td>
<td>39.5</td>
<td>60.5</td>
<td>78.3</td>
<td>22.0</td>
</tr>
<tr>
<td>LASSO</td>
<td>19.0</td>
<td>81.0</td>
<td>39.8</td>
<td>60.2</td>
<td>78.3</td>
<td>21.7</td>
</tr>
</tbody>
</table>

- **High TPR** (True Positive Rate) and **Low FDR** (False Discovery Rate) represents better performances.
- **INVASE** achieves better performances in comparison to all state-of-the-arts
Without Ground Truth – Prediction on Real Data

Mortality prediction performances on MAGGIC and PLCO datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Heart failure data</th>
<th>Cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAGGIC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3-year</td>
<td>5-year</td>
</tr>
<tr>
<td>Metrics</td>
<td>AUROC</td>
<td>AUPRC</td>
</tr>
<tr>
<td>INVASE</td>
<td>0.722</td>
<td>0.655</td>
</tr>
<tr>
<td>L2X</td>
<td>0.609</td>
<td>0.529</td>
</tr>
<tr>
<td>LIME</td>
<td>0.637</td>
<td>0.5596</td>
</tr>
<tr>
<td>Shapley</td>
<td>0.641</td>
<td>0.557</td>
</tr>
<tr>
<td>Knockoff</td>
<td>0.686</td>
<td>0.614</td>
</tr>
<tr>
<td>Tree</td>
<td>0.678</td>
<td>0.604</td>
</tr>
<tr>
<td>SCFS</td>
<td>0.683</td>
<td>0.623</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.692</td>
<td>0.615</td>
</tr>
</tbody>
</table>

- High AUROC and AUPRC represents better performances.
- INVASE achieves better performances in comparison to all state-of-the-arts
By-product – Prediction on Real Data

Mortality prediction performances on MAGGIC and PLCO datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>AUROC</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Labels</td>
<td>3 year</td>
<td>5 year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAGGIC</td>
<td>INVASE</td>
<td>.722±.005</td>
<td>.655±.010</td>
<td>.740±.005</td>
<td>.867±.006</td>
</tr>
<tr>
<td></td>
<td>Without INVASE</td>
<td>.720±.006</td>
<td>.639±.009</td>
<td>.730±.006</td>
<td>.855±.004</td>
</tr>
<tr>
<td></td>
<td>Labels</td>
<td>5 year</td>
<td>10 year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLCO</td>
<td>INVASE</td>
<td>.637±.007</td>
<td>.329±.013</td>
<td>.673±.007</td>
<td>.506±.006</td>
</tr>
<tr>
<td></td>
<td>Without INVASE</td>
<td>.629±.008</td>
<td>.324±.011</td>
<td>.657±.006</td>
<td>.485±.008</td>
</tr>
</tbody>
</table>

- Better instance-wise feature selection improves the prediction performances.
Qualitative Results on Real Data

- **Left**: Feature importance for each of 20 randomly selected patients in MAGGIC data
- **Right**: Average feature importance for different patient groups in MAGGIC data
Outline

• **Part I: Machine Learning Pipeline for Healthcare**
  • Data Imputation
    • Section 1: Generative Adversarial Imputation Network (GAIN)
    • Section 2: Multi-directional RNN (MRNN)
  • Model Interpretation
    • Section 3: Instance-wise Variable Selection (INVASE)
  • **Real-world Application**
    • Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)

• **Part II: Synthetic Data Generation**
  • Section 5: Time-series Generative Adversarial Nets (TimeGAN)
Section 4: Application - Machine Learning Pipeline

Raw dataset
- Temporal data (e.g., vital signs)
  - Temporal data
    - Interpolation
    - MRNN
- Static data (e.g., gender, age)
  - Static
    - Mean / Median
    - Masking
    - GAIN

Imputation

Model training
- RNN, LSTM, GRU
- Transformer
- Attention mechanism
- Temporal CNN

Trained predictive model

Main outputs
- Prediction results
- Performance metrics

Model interpretation
- INVASE

Explanations
- Feature importance
- Temporal importance
Application: One-shot Prediction for Mortality

- Admission
  - Static data (age, gender, etc.)
  - Admission time

- Prediction
  - Temporal data (vitals, lab tests, treatments, etc.)
  - Certain time point (y-hour)
  - End of time-series
  - Die or survive
  - ??-hours

Temporal data

Time
Application: Online Prediction for Ventilator

Admission

Static data
(age, gender, etc.)

Temporal data
(vitals, lab tests, treatments, etc.)

Prediction

Observation period

X-hours

Ventilator or not

Time

Admission time

Current (prediction) time

Current time + x hours

69
AutoMedic: Entire Pipeline of Machine Learning for Healthcare

Input interface (e.g., website)
- Load data
  - Temporal data (e.g., vital signs)
  - Static data (e.g., gender, age)
- Outlier detection
  - 95% or 99% Confidence interval
  - Manually input proper range
- Preprocess
  - Normalization
  - Encoding / Embedding
- Define problem
  - One-shot or online predictions
  - Observations define
  - Treatments define
  - Label define (e.g., mortality, ventilator)
- Defined data
  - Observations
  - Treatments
  - Labels

Imputation
- Temporal
  - Interpolation
  - Masking
  - MRRNN
  - T-GAIN
- Static
  - Mean / Median
  - GAIN
  - MICE

TrainTestSplit
- Train
- Val
- Test
- K-fold CV

Output interface (e.g., website)
- Interpretation
  - Feature importance
  - Temporal importance
  - Others
- Interpret
  - T-INVASE
  - Others
- Main outputs
  - Prediction results
  - Performance metrics
- Trained task-specific model
- Time-series model fit and predict
  - Time-series prediction models
    - RNN, LSTM, GRU
    - Transformer
    - Attention mechanism
  - Common options
    - Handle static features
    - Weighted loss
    - Handle irregular sampling
    - Masking
- Uncertainty
  - Prediction uncertainty
  - Confidence interval
  - Ensemble
  - Others

Output visualization

Feature selection
- Recursive feature selection
- Greedy feature selection
- None
Products: Web-based Clinical Decision Support

Step 2. Preprocess Dataset
- Filtering negative values.
- Min-max normalization selected.
- Encoding (stat): admit type.
- Preprocessing data... done.

Step 3. Define Problem
- Online problem (temp endpoint).
- Predict (temp): ventilator.
- Maximum sequence length: 24.
- Defining problem... done.

Step 4. Impute Dataset
- Median imputation (stat).
- Median imputation (temp).
- Imputing dataset... done.

Step 5. Fit and Predict
- Task selected: Classification.
- Method selected: LSTM Model.
- Uncertainty estimation: Ensemble.
- Interpretation method: INVAE.

Step 5. Fit and Predict

Prediction Task
- Choose a prediction task:
  - Classification
  - Regression

Prediction Method
- Choose a prediction method:
  - RNN Model
  - GRU Model
  - LSTM Model
  - Attention Model
  - TCN Model
  - Transformer Model

Uncertainty Estimation
- Choose an uncertainty estimation method:
  - Ensemble

Interpretation Method
- Choose an interpretation method:
  - INVAE

*Sensible hyperparameters will be selected for you.

☐ Commit
Visualizations of Results with AutoMedic

Accuracy

Predictions with uncertainty

Interpretation

Feature and temporal importance for patient ID: 1
Products: Jupyter-notebook Tutorials

Time-series AutoML framework - time-series prediction

ML-AIM (http://vanderschaar-lab.com/)

This notebook describes the user-guide of a time-series predictions application using Time-series AutoML framework. Time-series prediction is defined as following: utilize both static and temporal features to predict certain labels in the future. For instance, using the temporal data (vitals, lab tests) and static data (demographic information), we predict ‘whether the patient will die at the end of hospital stay’ or ‘whether the patient will get ventilator after 4 hours’.

- One-shot prediction: Predict the patient state at the end of the time-series at certain time point.
  - Example: Predict patient mortality (at the end of the hospital stays) after 24 hours from the admission.
- Rolling window (online) prediction:
  - Example: Predict ventilator after 24 hours from the current time point.

One-shot predictions for mortality

Online (Rolling window) predictions for ventilator

Admission  |  Prediction  |  Admission  |  Prediction
-----------|-------------|-------------|-------------
Static data (age, gender, etc.)  |  Temporal data (vitals, lab tests, treatments, etc.)  |  Static data (age, gender, etc.)  |  Temporal data (vitals, lab tests, treatments, etc.)

Die or Survive  |  X-hours  |  Observation period  |  X-hours

73
Products: Python Package

AutoMedic

Datasets

class datasets.CSVLoader(static_file=None, temporal_file=None) [source]

Load datasets from csv files.

- **static_file**
  file name of static data

- **temporal_file**
  file name of temporal data

load() [source]

Return both temporal and static datasets in PandasDataset format.

class datasets.PandasDataset(static_data, temporal_data) [source]

Return one-hot encoded dataset.

- **static_data**
  raw static data set.

- **temporal_data**
  raw temporal data set.

Preprocessing

class preprocessing.FilterNegative [source]
Outline

- Part I: Machine Learning Pipeline for Healthcare
  - Data Imputation
    - Section 1: Generative Adversarial Imputation Network (GAIN)
    - Section 2: Multi-directional RNN (MRNN)
  - Model Interpretation
    - Section 3: Instance-wise Variable Selection (INVASE)
  - Real-world Application
    - Section 4: End-to-End ML Pipeline for Medicine (AutoMedic)
- Part II: Synthetic Data Generation
  - Section 5: Time-series Generative Adversarial Nets (TimeGAN)
Section 5: Synthetic Data Generation - Motivation

Electronic Health Records → Privacy Regularization → Clinical Decision Supports

Machine Learning Community
Section 5: Synthetic Data Generation - Motivation

Impossible to share directly

Hospitals

Private medical data

Strong regularization

ML community

HIPAA/GDPR compliance

Medical data cannot be directly shared without strong regularization due to privacy and data protection regulations.
Section 5: Synthetic Data Generation

- Generate **synthetic data** which has **similar distribution** of the original data.
Section 5: Synthetic Data Generation

• **Enable to share** the private data (by sharing de-identified synthetic data) to machine learning community to develop machine learning tools easier.
Beyond Standard GAN Framework

• GANs achieves **tremendous success in static setting**.
• However, it did **not** achieve success in **time-series setting**.
• **Objective:** To generate time-series data with preserving temporal dynamics

• **Challenges:** Capture the distributions of features within each time point as well as complex dynamics of those variables across time points
Previous Works on Time-series Generation

- Autoregressive model
- Recurrent generative model

- We need a novel stochastic generative model that captures the distributions of features within each time point as well as complex dynamics of those variables across time points.
- We called this novel method as TimeGAN.

Problem Formulation: Time-series Generative Model

• Consider tuples of the form $(S, X_{1:T})$ with some joint distribution $p$, where
  – **Static features**: $S \in S$
  – **Temporal features**: $X \in X$

**Main Objective**

Given training data, learn a density $\hat{p}(S, X_{1:T})$ that best approximates $p(S, X_{1:T})$
Two Objectives of TimeGAN

Global Objective (sequence-level)
Matching the joint distribution

$$\min_{\hat{p}} D(p(S, X_{1:T}) \| \hat{p}(S, X_{1:T}))$$

It requires a perfect adversary.

Local Objective (stepwise)
Matching the conditional distribution

$$\min_{\hat{p}} D(p(X_t | S, X_{1:t-1}) \| \hat{p}(X_t | S, X_{1:t-1}))$$

It requires ground-truth sequence.
TimeGAN: **Encode**, Generate, and Iterate

- Provide mapping between the **feature space and latent space**, where the adversarial network learns the underlying **temporal dynamics** of the data

\[
\mathcal{L}_R = \mathbb{E}_{s, x_{1:T} \sim p} \left[ \| s - \tilde{s} \|_2 + \sum_t \| x_t - \tilde{x}_t \|_2 \right]
\]
TimeGAN: Encode, Generate, and Iterate

- The generator produces synthetic outputs through the latent space, and the discriminator operates on the basis of real vs. synthetic embeddings.

\[
\mathcal{L}_U = \mathbb{E}_{s,x_{1:T}\sim p} \left[ \log y_s + \sum_t \log y_t \right] + \mathbb{E}_{s,x_{1:T}\sim \hat{p}} \left[ \log (1 - \hat{y}_s) + \sum_t \log (1 - \hat{y}_t) \right]
\]
The generator receives sequences of embeddings of actual data to generate the next latent vector.

\[ \mathcal{L}_S = \mathbb{E}_{s, x_1:T \sim p} \left[ \sum_t \| h_t - g(x(S, h_{t-1}, z_t)) \|_2 \right] \]
TimeGAN: Jointly optimize

\[ \min_{\theta_e, \theta_r} (\lambda \mathcal{L}_S + \mathcal{L}_R) \]  
\[ \text{and} \quad \min_{\theta_g} (\eta \mathcal{L}_S + \max_{\theta_d} \mathcal{L}_U) \]
How to Evaluate Synthetic Data?

- **Desiderata 1:**
  - **Fidelity:** samples should be *indistinguishable* from real data
How to Evaluate Synthetic Data?

- **Desiderata 2:**
  - **Diversity:** samples should be *distributed to cover* that of real data
How to Evaluate Synthetic Data?

- **Desiderata 3:**
  - **Predictivity:** samples should be just as useful as real data when used for the same predictive purposes (i.e. train on synthetic, test on real)
Experiments - Desiderata 1: Fidelity

Table: Results on Multiple Time-Series Datasets (Bold indicates best).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Sines</th>
<th>Stocks</th>
<th>Energy</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminative Score</td>
<td>TimeGAN</td>
<td>0.011±0.008</td>
<td>0.102±0.021</td>
<td>0.236±0.012</td>
<td>0.161±0.018</td>
</tr>
<tr>
<td></td>
<td>RCGAN</td>
<td>0.022±0.008</td>
<td>0.196±0.027</td>
<td>0.336±0.017</td>
<td>0.380±0.021</td>
</tr>
<tr>
<td></td>
<td>C-RNN-GAN</td>
<td>0.229±0.040</td>
<td>0.399±0.028</td>
<td>0.499±0.001</td>
<td>0.462±0.011</td>
</tr>
<tr>
<td></td>
<td>T-Forcing</td>
<td>0.495±0.001</td>
<td>0.226±0.035</td>
<td>0.483±0.004</td>
<td>0.387±0.012</td>
</tr>
<tr>
<td></td>
<td>P-Forcing</td>
<td>0.430±0.027</td>
<td>0.257±0.026</td>
<td>0.412±0.006</td>
<td>0.489±0.001</td>
</tr>
<tr>
<td></td>
<td>WaveNet</td>
<td>0.158±0.011</td>
<td>0.232±0.028</td>
<td>0.397±0.010</td>
<td>0.385±0.025</td>
</tr>
<tr>
<td></td>
<td>WaveGAN</td>
<td>0.277±0.013</td>
<td>0.217±0.022</td>
<td>0.363±0.012</td>
<td>0.357±0.017</td>
</tr>
</tbody>
</table>

- Samples should be indistinguishable from real data
Experiments - Desiderata 2: Diversity

- Samples should be distributed to cover that of real data

(a) TimeGAN
Experiments - Desiderata 3: Predictivity

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Sines</th>
<th>Stocks</th>
<th>Energy</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive</td>
<td>TimeGAN</td>
<td>.093±.019</td>
<td>.038±.001</td>
<td>.273±.004</td>
<td>.303±.006</td>
</tr>
<tr>
<td>Score</td>
<td>RCGAN</td>
<td>.097±.001</td>
<td>.040±.001</td>
<td>.292±.005</td>
<td>.345±.010</td>
</tr>
<tr>
<td></td>
<td>C-RNN-GAN</td>
<td>.127±.004</td>
<td>.038±.000</td>
<td>.483±.005</td>
<td>.360±.010</td>
</tr>
<tr>
<td></td>
<td>T-Forcing</td>
<td>.150±.022</td>
<td>.038±.001</td>
<td>.315±.005</td>
<td>.310±.003</td>
</tr>
<tr>
<td></td>
<td>P-Forcing</td>
<td>.116±.004</td>
<td>.043±.001</td>
<td>.303±.006</td>
<td>.320±.008</td>
</tr>
<tr>
<td></td>
<td>WaveNet</td>
<td>.117±.008</td>
<td>.042±.001</td>
<td>.311±.005</td>
<td>.333±.004</td>
</tr>
<tr>
<td></td>
<td>WaveGAN</td>
<td>.134±.013</td>
<td>.041±.001</td>
<td>.307±.007</td>
<td>.324±.006</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>.094±.001</td>
<td>.036±.001</td>
<td>.250±.003</td>
<td>.293±.000</td>
</tr>
</tbody>
</table>

Table: Results on Multiple Time-Series Datasets (Bold indicates best).

- Samples should be **just as useful as** real data when used for the same predictive purposes (i.e. train on synthetic, test on real)
Conclusion: Machine Learning for Healthcare

Objective

Clinical decision supports

ML pipeline for healthcare

Model Interpretation

Prediction

Data Imputation → Model Training

Synthetic data generation

ML tools

Hospitals → Synthetic data generation

Deidentified

Private medical data → Synthetic medical data → ML community
Conclusion: Machine Learning for Healthcare

- Genetic disease prediction
- Optimal treatment recommendation
- Optimal screening
- Early warning systems
- Resource allocation
References – Publications in this Talk

• Data Imputation:

• Model Interpretation:

• Time-series Generation:
References – Other Publications

• Privacy:

• Active Sensing:
• **Adversarial Learning:**
  
  
  

• **Clinical Publications:**
  
  
Acknowledgements – Clinical Collaborators

• **Intensive Care Units:**
  – Dr. Scott Hu¹, Dr. Ari Ercole², Dr. Tomas Daniels ⁵

• **Cardiology:**
  – Dr. Martin Cadeiras¹, Professor Raffaele Bugiardini³, Professor Paolo Puddu⁴

• **Patient Subgroup Analysis:**
  – Dr. Camelia Davtyan¹, Dr. Mindy Ross¹

• **Synthetic Medical Data Generation:**
  – Dr. Lydia Drumright²

• **Affiliations**
  – ¹University of California, Los Angeles, USA
  – ²University of Cambridge, UK
  – ³University of Bologna, Italy
  – ⁴Sapienza University of Rome, Italy
  – ⁵University Hospital Southampton, UK
Acknowledgements – Machine Learning Researchers

• **Data Imputation:**
  – Professor William R. Zame\(^1\), James Jordon\(^2\)

• **Model Interpretation:**
  – James Jordon\(^2\)

• **AutoMedic:**
  – Daniel Jarrett\(^3\), Ioana Bica\(^2\), Zhaozhi Qian\(^3\), Dr. Ari Ercole\(^3\)

• **Time-series Generation:**
  – Daniel Jarrett\(^3\)

• **Affiliations**
  – \(^1\)University of California, Los Angeles
  – \(^2\)University of Oxford
  – \(^3\)University of Cambridge
End-to-End Machine Learning Frameworks for Medicine: Imputation, Interpretation and Synthetic Data Generation

Discussion Session
(Question & Answer)

Jinsung Yoon

Advisor: Prof. Mihaela van der Schaar

April 13th, 2020