Track patient recovery in real-time by processing streaming data

BIOMEDICAL DATA DESIGN

TA: Haoyin Xu

Group:Zhenyu Xiao
Haobin Zhou
Yimeng Xu
Emma Cardenas

01
Reprocess the data

Reprocess the data

```python
def extract_data_optional(
    patient_batch,
    patient_offset,
    data,
    time_length=1,
    max_min_var=False
):
    """
    Summary: align data and interpolate missing values

    Args:
        patient_batch: the list of wanted patient id, used to split data
        patient_offset: the dataframe of patient offset data, including patientunitstayid, unitdischargeoffset
        data: the dataframe of data, including patientunitstayid, observationoffset, value
        kernel: the self-defined kernel function for Gaussian Process Regressor
        align: decide whether align the data
        time_length: decide how long we sample
        max_min_var: decide whether we calculate the max, min and the var

    Returns:
        data_full: the dataframe of aligned and interpolated data, including patientunitstayid, observationoffset, value
        data_full_index: the series of the index of the first occurrence of each patient
    """
```

```python
def flatten(df):
    """
    Summary: Function to flatten the dataframe.
    Args:
        df: the dataframe of data, including patientunitstayid, observationoffset, and features

    Returns:
        df_flat: the dataframe of flattened data
    """
    df = df.set_index(['patientunitstayid', 'observationoffset'])
    df_flat = df.unstack(level='observationoffset')
    df_flat.columns = ['{}_hour{}'.format(*col) for col in df_flat.columns]
    df_flat = df_flat.reset_index()
    return df_flat
```
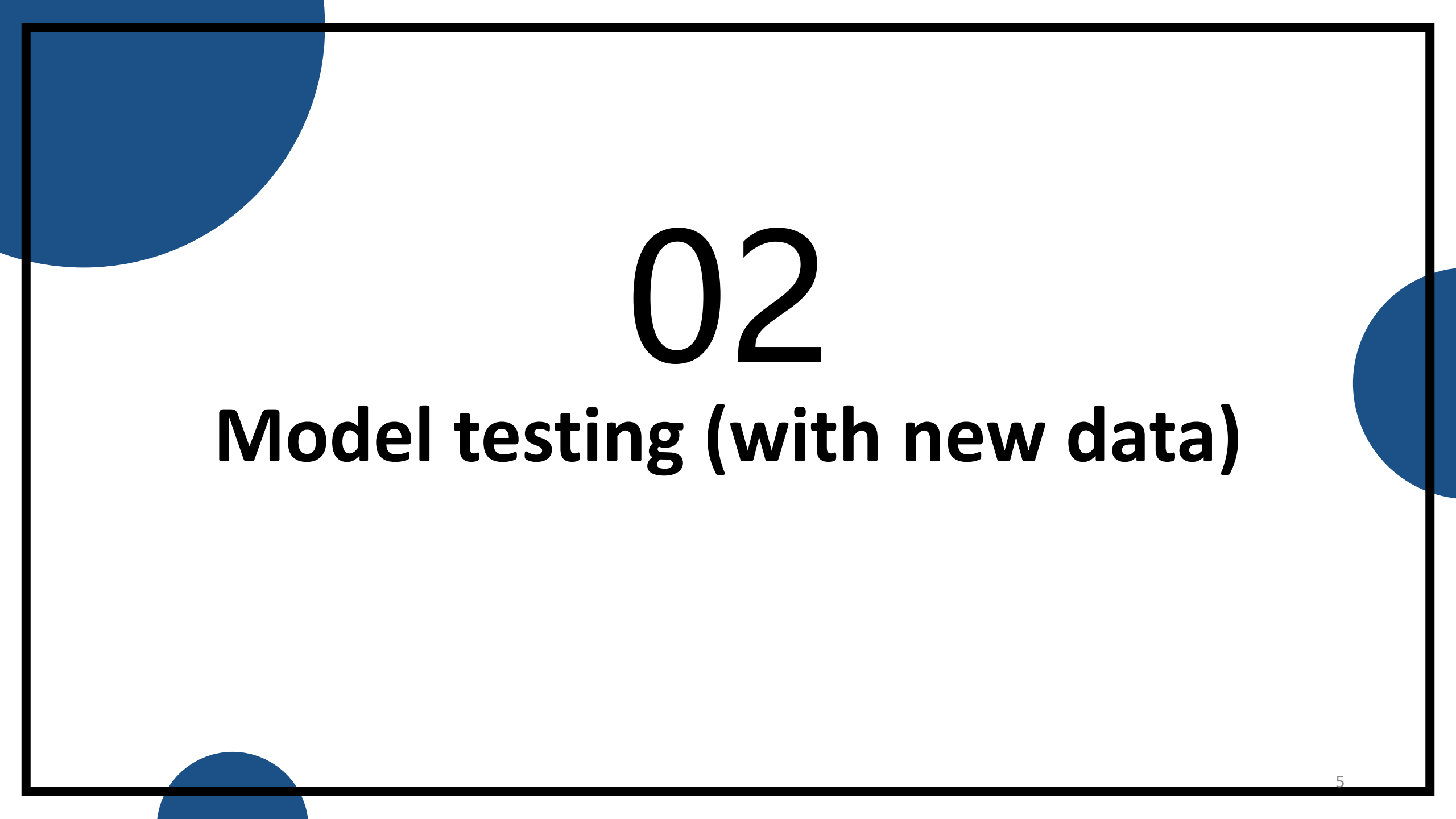
1.Use the three highest frequency features and change intervals

2.Flatten these features

Reprocess the data

patientunits	heartrate_hour0	heartrate_hour1	heartrate_hour2	heartrate_hour3	heartrate_max_hour0
141266	95.3164557	97.775	100.3037975	102.9285714	102
141288	101.5578947	90.1875	84.56666667	91.32051282	105
141436	76.96	69.16049383	77.22352941	95.57894737	90
141462	86.29896907	86.18666667	82.50632911	92.98924731	96
141515	87.43157895	103.1034483	85.3814433	82.87912088	94
141751	107.4367816	105.6067416	112.0120482	106.1702128	137
141822	74.44329897	79.28125	68.35416667	77.69892473	86
142173	74.10465116	76.3902439	81.36708861	75.81818182	86
142483	77.10810811	90.07594937	75.20833333	77.27083333	100
142636	78.44642857	69.61458333	69.14285714	67.92682927	132
142858	62.71212121	64.31506849	52.5	52.58695652	82
142923	77.04225352	72.27380952	78.33333333	110.0625	84
143025	68.63855422	68.33333333	69.69135802	71.70512821	74
143101	90.08641975	93.69879518	92.1	97.21794872	101
143103	73.34545455	80.89411765	90.42857143	81.0989011	82
143265	85.63461538	80.42045455	67.15384615	64.76923077	90
143400	79.55789474	74.2247191	74.83544304	67.625	90
143469	95.37704918	96.07407407	95.81818182	98.61538462	106
143578	62.83050847	53.85714286	53.09333333	53.21052632	70
143638	58.38947368	64.3908046	67.8974359	72.65909091	64
143642	89.11538462	92.20253165	93.34883721	99.1375	96
143860	72.30379747	69.05128205	72.18181818	75.06329114	80
143885	63.18947368	58.86075949	59.58227848	60.02564103	84
143994	124.26	110.6860465	95.62820513	102.0384615	166
144034	89.41025641	87.74358974	90.05128205	108.1044776	106
144108	72.86538462	75.97468354	74.85897436	75.16666667	80
144252	81.25	90.7254902	74.74545455	80.98876404	90
144409	77.73255814	69.21518987	77.05128205	72.02564103	94
144634	69.23684211	62.53521127	60.26470588	80	75
144722	84.275	89.01204819	86.55813953	89.39240506	102
145183	74.45348837	68.68674699	69.35897436	71.2278481	101
145270	101.7413793	100.3225806	96.84337349	99.01265823	138

Raw data: Without interpolation

02

Model testing (with new data)

Model testing

Using 13 features:

```
accuracy: 72.01%
confusion_matrix:
[[4586 2180]
 [1674 5331]]
```

Not Flatten，Max & Min & Var

```
Accuracy: 0.6326724910394266
Confusion matrix:
[[5356 3572]
 [2987 5941]]
Classification report:
              precision    recall  f1-score   support

         0.0       0.64      0.60      0.62      8928
         1.0       0.62      0.67      0.64      8928

    accuracy                           0.63     17856
   macro avg       0.63      0.63      0.63     17856
weighted avg       0.63      0.63      0.63     17856
```

Not Flatten, Without max…

```
Accuracy: 0.6177755376344086
Confusion matrix:
[[5335 3593]
 [3232 5696]]
Classification report:
              precision    recall  f1-score   support

         0.0       0.62      0.60      0.61      8928
         1.0       0.61      0.64      0.63      8928

    accuracy                           0.62     17856
   macro avg       0.62      0.62      0.62     17856
weighted avg       0.62      0.62      0.62     17856
```

Flatten，Max & Min & Var

```
Accuracy: 0.6727150537634409
Confusion matrix:
[[1424  808]
 [ 653 1579]]
Classification report:
              precision    recall  f1-score   support

           0       0.69      0.64      0.66      2232
           1       0.66      0.71      0.68      2232

    accuracy                           0.67      4464
   macro avg       0.67      0.67      0.67      4464
weighted avg       0.67      0.67      0.67      4464
```

Flatten, Without max…

```
Accuracy: 0.6236559139784946
Confusion matrix:
[[1345  887]
 [ 793 1439]]
Classification report:
              precision    recall  f1-score   support

           0       0.63      0.60      0.62      2232
           1       0.62      0.64      0.63      2232

    accuracy                           0.62      4464
   macro avg       0.62      0.62      0.62      4464
weighted avg       0.62      0.62      0.62      4464
```

Thank you