一、数据处理部分：

首先是基本的数据清理：

```
1、处理空值：有空缺的行比较少，所以全部去除；
2、处理异常值：数据集中有一些不符合实际的极端数值，结合实际情况，我们去除了呼吸率超过60,氧饱
3、统一序列长度：数据集中序列长度不固定，考虑到大多数序列长度并没有24，我们将序列长度统一为2
    3.1 处理方法为：假如原本序列长度超过20，那么只保留最后的20个时刻；假如原本序列长度不足2
```

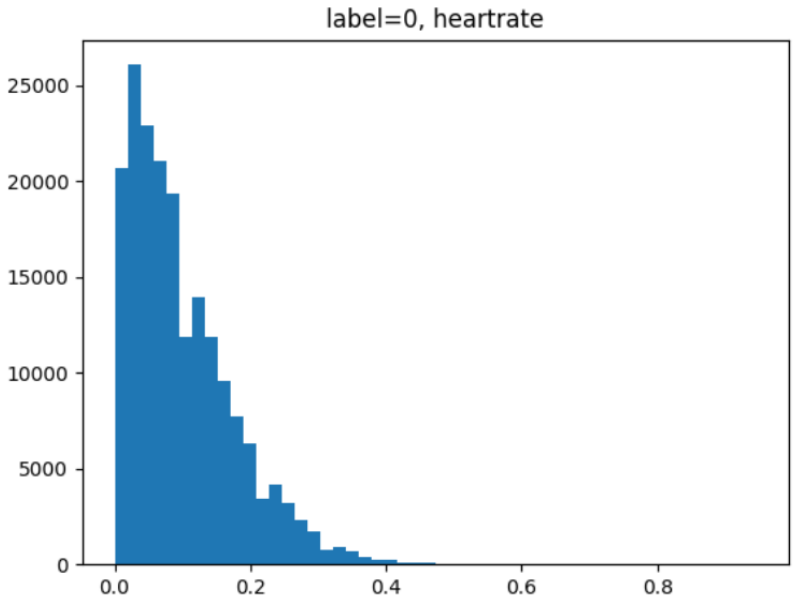
然后是简单的扩充和规范化：

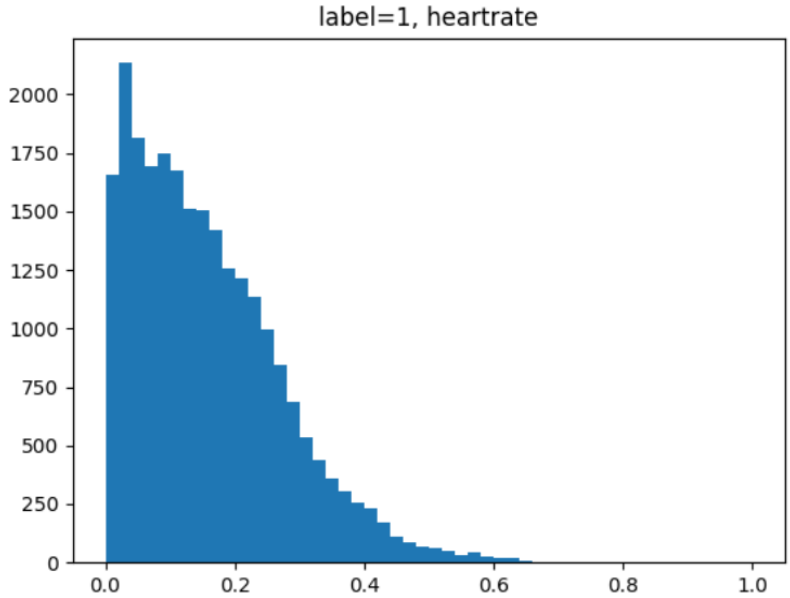
对于四个字段heartrate,resprate,map,o2sat中的每一个，都进行如下处理(下面用attr表示前述任意字段)：

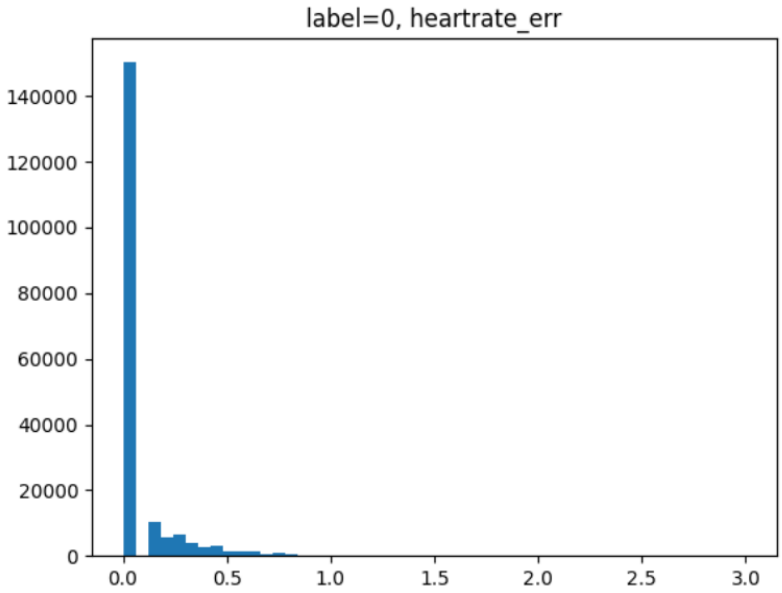
```
1、加入新字段attr_DELTA表示差分后的结果：出于两个原因：1、体征稳定的病人更加安全；2、病人本
2、加入新字段attr_ERR表示异常程度：为了利用说明文档给出的正常值范围，新增该字段表示属性偏离
    2.1 注意到许多时候病人指标会在正常范围边缘(比如氧饱和度在94~96之间波动而正常边界是95),
    2.2 根据背景知识，氧饱和度指标是越高越好(100是最好的)。
3、将attr从原始数据改为数据离正常范围中间值的距离。这是因为原始数据太小和太大(比如心率过低和
4、规范化：缩小数据范围到[0,1]或者[0,10]，确保能正常学习。
```

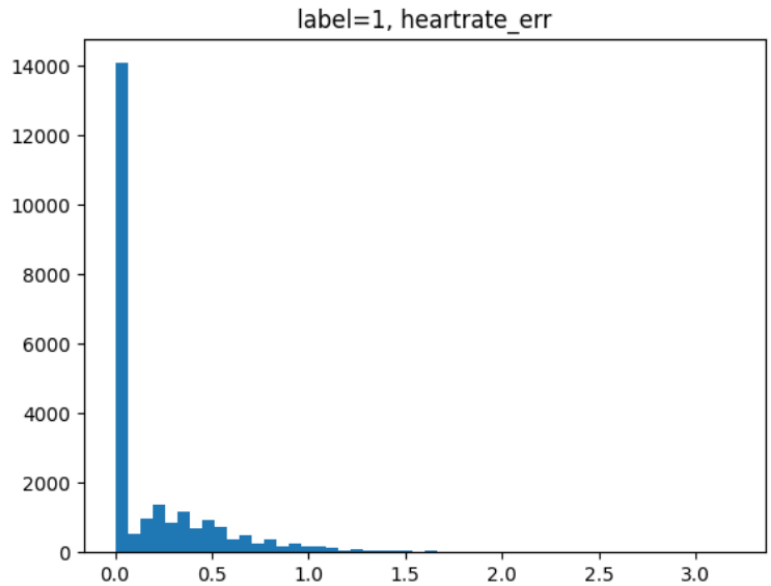
处理后的部分字段分布图：

左右图分别表示标签为0和1的训练集数据。在预处理完后，数据大小和生存情况的关系更直观。


label=0, heartrate
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8


label=1, heartrate
2000
1750
1500
1250
1000
750
500
250
0
0.0
0.2
0.4
0.6
0.8
1.0


label=0, heartrate_err
140000
120000
100000
80000
60000
40000
20000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0


label=1, heartrate_err
14000
12000
10000
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0