项目编号：S202210698280

西安交通大学

大学生创新训练项目

结题报告书

项目名称：基于多元数据分析的重症监护室病人健康状态预警方法研究

起止年月：2022 年 5 月至 2023 年 5 月

负 责 人：喻彭

手　　机：13036784398

邮　　箱：yu.peng@stu.xjtu.edu.cn

项目成员：吴天阳，王承杰，储文煜

指导老师：孙剑

批准经费：5000 已用金额：705.59

填报日期：2023 年 5 月 22 日

报告题目：基于多元数据分析的重症监护室病人健康状态预警方法研究

学生姓名：喻彭，吴天阳，王承杰，储文煜

指导老师：孙剑

中文摘要：医疗实践活动产生海量的数据，随着信息化的发展，我们得以及时记录医疗数据，然而医疗重症监护数据是一个没有显著规律的动态系统, 数据随时间波动复杂变化，而且存在数据稀疏性强、不规则程度高等问题. 因此数据以往没有得到充分利用于临床诊断. 我们以脓毒症（可导致 SIC 、DIC）为例，获取了重症监护室记录 MIMIC-IV 数据集，对其进行筛选、异常值排查、数据缺失处理、实时状态标定等预处理，使用随机梯度下降优化的线性模型（SGD）、SVM 模型、决策树、随机森林、K 近邻等机器学习方法进行模型构建，用 K-折交叉验证准确率、测试集准确率和 AUC 等指标分析相关结果. 对于 DIC，我们发现在序列长度为 2、时间段长度为 4h 时，模型准确率较高；对于 SIC 综合考虑，我们发现在序列长度为 3、时间段长度为 8h 时，模型效果较好. 最后用 SHAP 方法进行解释性分析，分别得到整体数据集的可解释性分析，并发现由 SHAP 给出的重要度与随机森林模型给出的重要度排名几乎相同，说明 SHAP 确实能够对模型进行有效的数据解释；进一步，我们从标签对应的数据集中各取出一个数据进行具体可解释性分析，并根据结果对患者的治疗方向提出建议.

英文摘要：Medical practice generates massive data. With the development of information technology, we can record medical data in real-time. However, ICU data is a dynamic system without significant patterns and has problems such as sparsity and irregularity. Therefore, it has not been fully utilized for clinical diagnosis. We used the MIMIC-IV dataset to build models using machine learning methods and analyzed the results with accuracy and AUC. We found that the model accuracy was higher for DIC when the sequence length was 2 and the time period was 4h, and for SIC when the sequence length was 3 and the time period was 8h. Finally, we used SHAP for interpretability analysis.

关键词：医疗数据分析，机器学习方法，可解释性分析.

目录

1 绪论

医疗实践活动产生海量的数据, 随着信息化的发展, 我们得以及时记录医疗数据, 深度挖掘和利用这些数据对提高医疗、护理质量以及患者安全都具有重大意义. 本项目拟以脓毒症为例，收集重症监护室记录数据，对其进行筛选、异常值排查、数据缺失处理、标记等预处理，设计多种涵盖机器学习方法进行模型构建，分析相关结果并进行验证. 搭建实时的检测预警系统，预测患者患病风险，为医生诊断提供有效参考.

近些年来，机器学习对脓毒症的早期预测有了极大的进展. 2018 年，Nemati 等根据纳入患者实时监测的 EMR 数据中提取的 65 个变量，机器学习构建出了人工智能脓毒症专家（artificial intelligence sepsis expert，AISE）预警模型，可以优先临床 4-12 h 预测脓毒症的发生. 2020 年，Burdick 等根据 270 438 例住院或急诊成人患者的心率、呼吸频率、收缩压、舒张压、体温、血氧饱和度，利用梯度增强方法所开发的预警模型就可以提前 48 h 预测脓毒症的发生. 2020 年，Yang 等利用极端梯度增强（eXtreme Gradient Boosting，XGBoost）算法提取了 34 285 例成人重症患者 EHR 数据中的 168 个临床特征变量, 并训练出了一个可解释的人工智能脓毒症预测模型（explainable AI sepsispredictor，EASP），该模型能够提前预测脓毒症患病风险.

通过机器学习继续提高脓毒症早期预测的时间和精确性是非常值得研究的问题，将会降低脓毒症的治疗成本改善结果，并为医疗系统、医务人员和患者带来极大的益处.

本文将基于 MIMIC-IV 2.2 数据集对 299 712 位病人的住院及化验信息，提取与脓毒症判断相关的信息，并对数据进行插补、异常值矫正、缺失值处理后，使用不同的机器学习模型对病人是否患有 DIC 进行预测，使用 K-折交叉验证并用 ROC 对模型进行评估，从而得到最优的模型与数据集处理方法，最后使用 SHAP 值对模型的可解释性进行分析，以解释不同的特征对最终预测结果的影响.

2 数据集分析及预处理

2.1 数据集介绍

我们已在 PhysioNet 上获取了 MIMIC 数据集使用权限，使用的为当前最新的 MIMIC-IV 2.2 数据集 [1, 2]，该数据集在之前的基础上修复了部分错误，并加入了护理信息.

MIMIC-IV 数据集通过医院中的数字电子健康记录系统 (Electronic Health Recor system, EHR) 对全部入院病人信息进行记录，通过 BIDMC 的 MetaVision(iMDSoft) 系统对 ICU 病人的生命体征数据进行记录，出于对患者隐私的保护，全部患者姓名与入院年份均经过 Hash 编码，通过对数据的大致分析，我们得到入院病人总数为 299712 人，住院病人 180733 人，ICU 病人 50920 人. 图1中分别显示了入院与进入 ICU 的年龄占比分布，并通过分析得到该数据集中进入医院的病人中死亡率为 9.70%，而进入 ICU 病人的死亡率为 32.36%，基本符合实际情况.

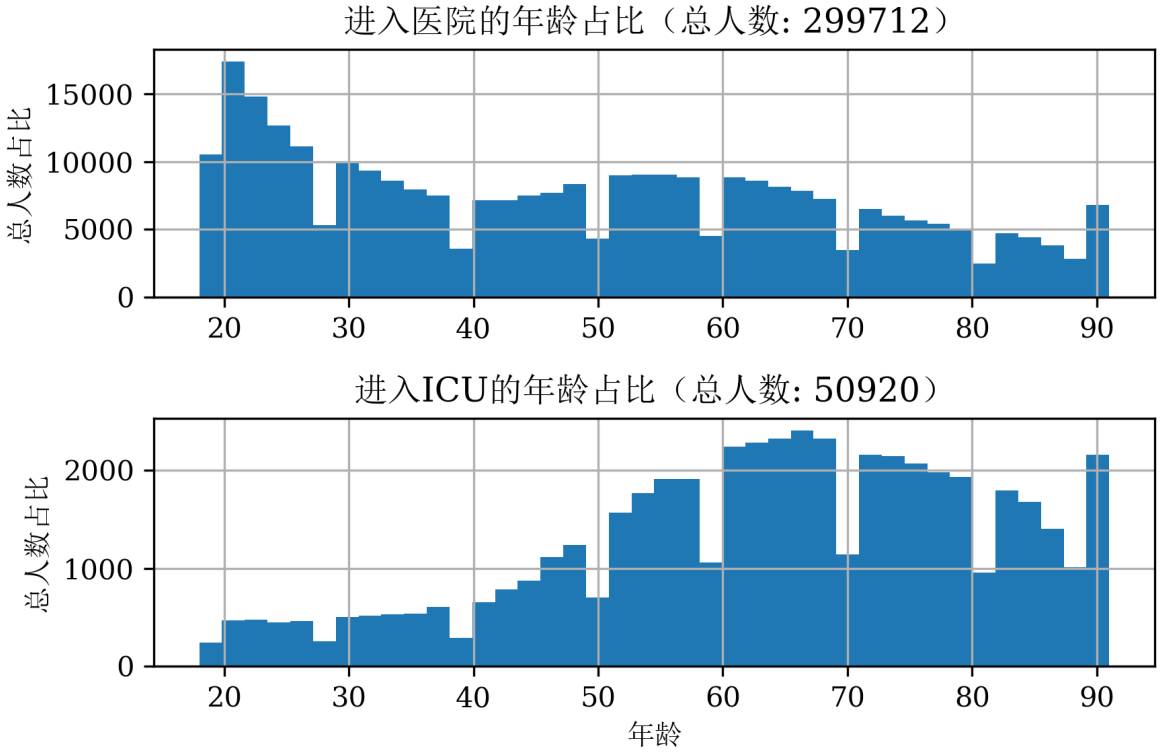


图 1: 数据集中年龄占比分布

2.2 数据预处理

2.2.1 数据提取

由于 MIMIC-IV 中的数据都是每次的化验结果，我们首先要找到与 DIC 症状相关的化验信息，并将所有相关化验信息整合在一起，基于文献 [3] 中给出的信息，我们将表1中的 8 个化验信息作为判断脓毒症的特征. 其中前 5 项与 DIC 的判定有关，后 3 项为常见化验项目，用于发现与 DIC 可能存在的潜在关系.

经过分析处理后，我们得到如图2，两个图分别显示了每种标签的化验条例数目和每个时刻下的检测数据中所包含的非空条例数目. 从图中不难看出，D-Dim 的条例数目最少，很有可能是因为该化验只会对感染某些与呼吸道相关疾病时才会检验，从第二个

图中可以看出大多数的时刻下的数据都只包含 3 个特征数据，所以我们首先需要按照一个时间段对数据进行插补.

数据全称	缩写	MIMIC-IV 数据编号
1. 血小板计数 (Platelet Count)	PLT	227457
2. 凝血酶原时间 (Prothrombin time)	PT	227465
3. 凝血酶原时间的国际标准化比值 (International Normalized Ratio)	INR	227467
4. D-二聚体 (D-Dimer)	D-Dimer	225636
5. 纤维蛋白原 (Fibrinogen)	FIB	227468
6. 二氧化碳分压 (Venous CO2 Pressure)	pCO2	226062（动脉）
7. 酸碱度 (pH)	pH	223830（动脉）
8. 氧分压 (Venous O2 Pressure)	pO2	226063（动脉）

表 1: 相关数据对应缩写及编号

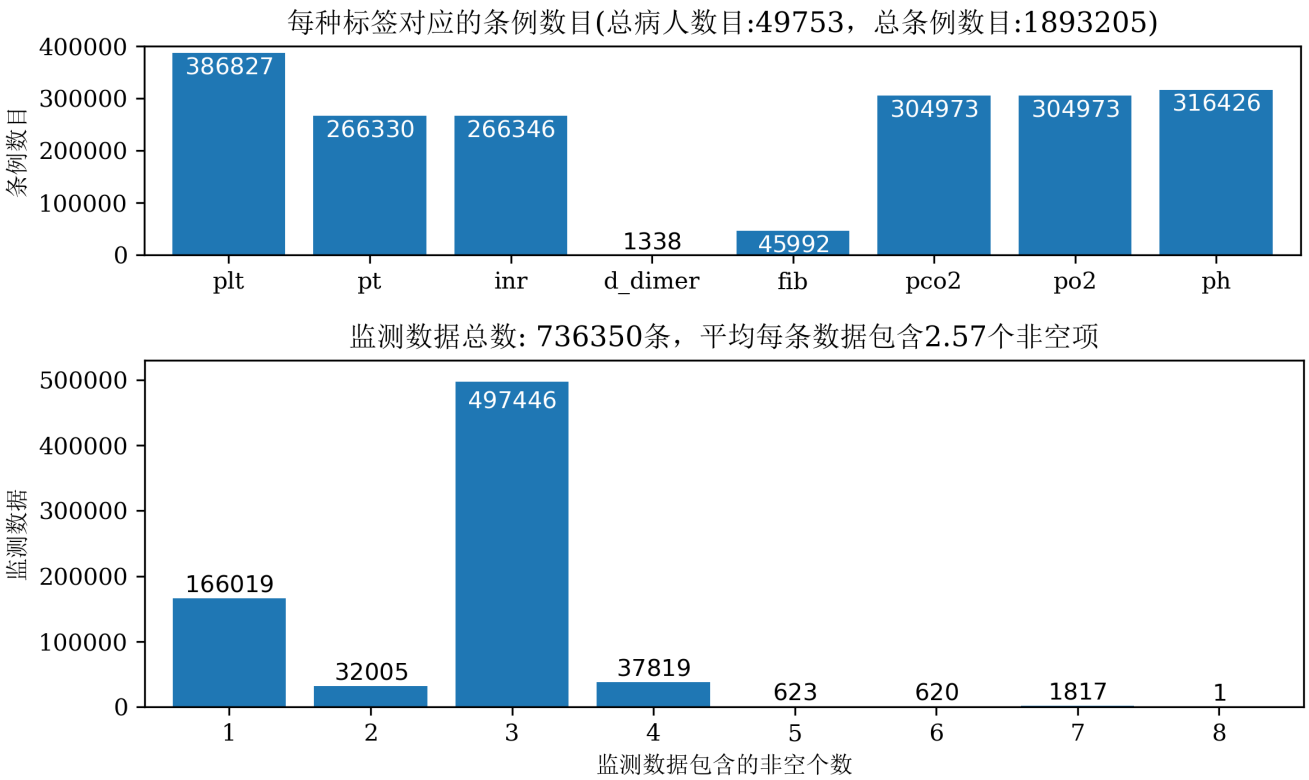


图 2: 数据提取结果

2.2.2 数据插补

进行数据插补前，我们首先要以某个起始时刻 T_0 作为基准时间，其他时刻均对 T_0 做差，并转化为小时为单位的时间，具体实现如算法1所示.

Algorithm 1 时间转换

```
charttimes = df['charttime']
# 以最小时间戳 charttimes.min() 作为基准时间，计算其他时间与其的相对时间差
time_format = '%Y-%m-%d %H:%M:%S'
base_time = datetime.strptime(charttimes.min(), time_format)

# 计算与基准时间的相对时间差（单位：小时，浮点型）
def relative_hours(date_string):
    date = datetime.strptime(date_string, time_format)
    relative_days = (date - base_time).days  # 相对天数
    relative_seconds = (date - base_time).seconds  # 相对秒数
    relative_hours = 24 * relative_days + relative_seconds / 3600
    return relative_hours
```

数据插补的方法就是将同一个病人在同一时间段内的信息进行合并，从而得到一个时间段内的数据整体作为新的数据集. 我们记时间段长度为 L 小时，例如当 $L=8$ 时，说明需要将两个相邻时间小于等于 8 小时的数据进行合并.

设当前病人具有 N 条信息，将第 $i\in\{1,\cdots,N\}$ 条信息记为 $\boldsymbol{x}^{(i)}=\{x_k^{(i)}\}_{k=1}^8\in\mathbb{R}^8$，该信息的相对时间为 t_i；设第 j 个时间段为 $[T_j,T_j+8)$，其对应的特征向量为 $\boldsymbol{y}^{(i)}=\{y_k^{(i)}\}_{k=1}^8\in\mathbb{R}^8$. 我们设计了以下两种插补方法：

覆盖插补 使用该时间段中的最后一次非空特征值做插补：

$$y_k^{(j)}\leftarrow x_k^{(i)},\quad s.t.\begin{cases}t_i=\max\limits_{t_i\in[T_j,T_j+L)}t_i,\\ x_k^{(i)}\text{ 非空.}\end{cases}\tag{2.1}$$

均值插补 若该时间段中存在对于第 $k\in[1,8]$ 个特征有多条信息，则取均值做插补：

$$y_k^{(i)}\leftarrow\frac{1}{N_k^{(i)}}\sum_{\substack{t_i\in[T_j,T_j+8)\\ x_k^{(i)}\text{ 非空}}}x_k^{(j)}\tag{2.2}$$

其中 $N_k^{(i)}=\#\{t_i:t_i\in[T_j,T_j+8),x_k^{(i)}\text{ 非空}\}$，$\#S$ 表示集合 S 的基数.

符号	意义（以下定义中的信息均来自同一个病人）
$x_{time}^{(i)}$	**时间戳：** 第 i 个信息 $\boldsymbol{x}^{(i)}$ 对应的时刻
$S=(\boldsymbol{x}^{(i_1)},\cdots,\boldsymbol{x}^{(i_k)})$	**信息段：** 按照信息的时间戳从小大大排序，并且满足 $x_{time}^{(i_k)}-x_{time}^{(i_1)}\leqslant L$
$S_1\prec S_2$	**信息段的序关系：** $\boldsymbol{x}_{time}<\boldsymbol{y}_{time},(\boldsymbol{x}\in S_1,\boldsymbol{y}\in S_2)$
$S_1\sim S_2$	**连续信息段：** $\min\limits_{\boldsymbol{y}\in S_2}y_{time}-\max\limits_{\boldsymbol{x}\in S_1}x_{time}\leqslant L,(S_1\prec S_2)$
$S_1\sim S_2\sim\cdots\sim S_k$	**多个连续信息段：** $S_i\sim S_{i+1},(i=1,\cdots,k-1)$
$B=\{S_1,S_2\cdots,S_k\}$	**信息块：** $S_1\sim\cdots\sim S_k$ 且 $k\geqslant 2$

表 2: 符号定义

为更清晰的描述插补后的数据，我们给出如表2中对信息段与信息块的定义，上述定义说明：同一病人的任意两个信息段 $S_i, S_j, (i \neq j)$ 均有序关系成立，并且信息块 B 是由至少两个连续信息段构成的.

我们之所以要这样定义信息块，是因为首先要保证其中的信息段之间的时间差不超过 L 小时，这样才能使得新的数据信息具有时序性，并且要求信息段至少有两个，因为我们希望通过前几个连续的时间段预测未来的一个时间段是否会患有 DIC.

数据插补得到的最终结果为每个病人的全部信息块，经过插补操作后每种标签对应的非空信息占比如图3

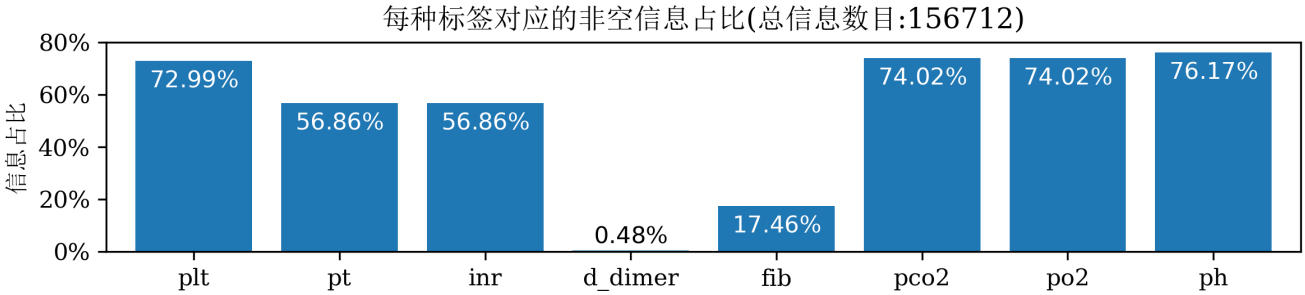


图 3: 插补操作后非空信息占比

2.2.3 异常值矫正

我们按照表3中定义的极限范围对数据进行异常值矫正，即若数据超过极大范围则调整至最大极限值，反之亦然.

名称	缩写	正常范围	极限范围	单位
血小板计数	`plt`	$[100, 300]$	$[10, 300]$	$\times 10^9$/L
凝血原酶时间	`pt`	$\leqslant 16$	$[7, 30]$	s（秒）
国际标准化比值	`inr`	$[0.8, 1.4]$	$[0.5, 2]$	无
D-二聚体	`d_dimer`	500 左右	$[100, 10000]$	ug/L
纤维蛋白原	`fib`	$[1000, 2000]$	$[10, 2000]$	mg/L
二氧化碳分压	`pco2`	$[35, 45]$	$[5, 250]$	mmHg
氧分压	`po2`	$[80, 100]$	$[20, 500]$	mmHg
酸碱度	`ph`	$[7.35, 7.45]$	$[6, 8]$	无

表 3: 每类数据的极限范围及对应单位

2.2.4 数据填补

从图3中可以看出，经过插补后的数据中 `fib` 偏少，而 `d_dimer` 极少，我们尝试了一下三种方法对缺失数据进行填补.

均值填补 若一个信息块 B 中的某个信息段 $s \in B$ 的第 k 维属性值为空，分以下两种情况：

- 若存在该信息块中其他信息段该属性值非空，则使用该属性的其他非空值的均值进行填补：

$$s_k \leftarrow \frac{1}{\#\{s : s \in B, s_k \text{ 非空}\}} \sum_{s \in B, s_k \text{ 非空}} s_k$$

- 若该信息块中 k 维属性值全部为空，则直接用正常值进行填补，每种属性的正常值取表3中正常范围的中位数.

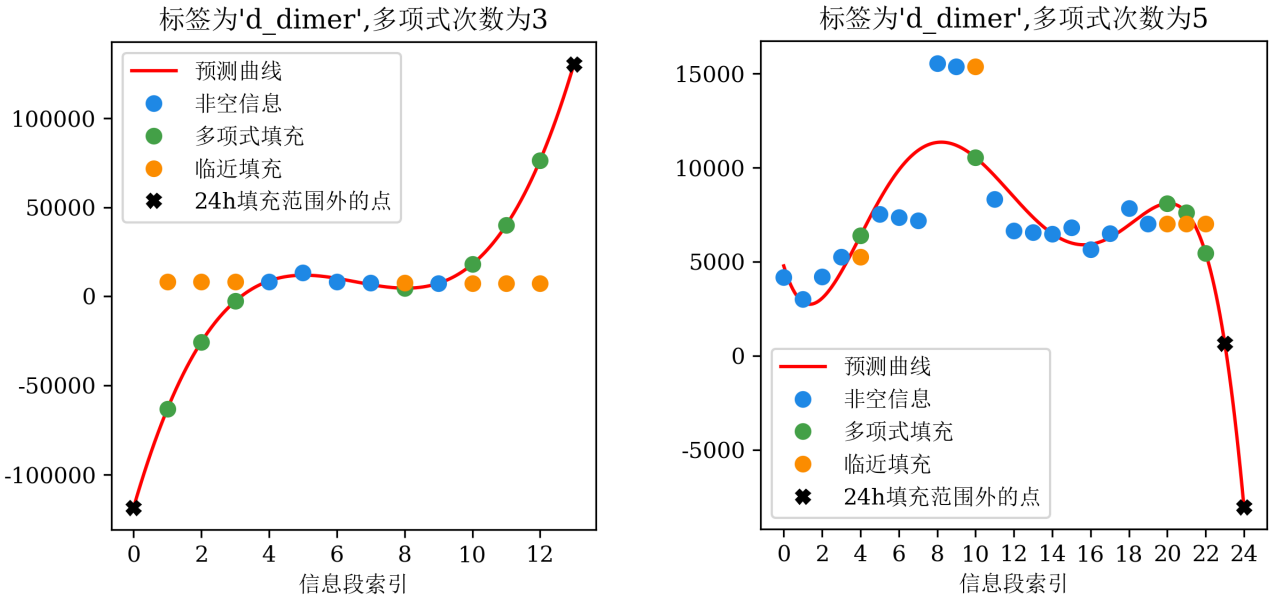
自适应多项式填补策略 考虑到一个信息块中的各项属性之在每个时间段下不会是恒定值，于是我们在以上策略的基础上引入自适应多项式填补策略，假设信息块 S 中第 k 维属性值有 m 个非空值，分为一下三种情况：

- $m = 0$，全部为空值，则全部使用正常值进行填补.
- $m = 1$，仅有一个数据，则全部使用该数据进行填充.
- $m > 1$，使用 $[m/2]$ 阶多项式对空值进行回归预测.（$[x]$ 表示对实数 x 向下取整）

假设回归数据中 x 为对应的非空值，先进行多项式特征提取转化为线性回归问题（若多项式阶数 $n = 3$，则提取出 $x^3, x^2, x^1, 1$ 项），然后进行标准化处理，再使用 SVD 分解求解最小二乘回归问题，最后将多项式模型对空值的索引进行插值，得到对应的预测结果.

基于 24 小时有效范围的临近填补 考虑到一个信息块 B 中的内容如果使用较长时间差的点进行填补则会出现较大的误差，并且如果使用高阶多项式进行拟合可能出现 Runge 现象（如图4所示），所以我们基于上面两个方法提出一种简单易行的且更符合现实的填补方法：

考虑使用 24 小时内的非空信息对空信息进行填补，对于一个信息块中空信息，我们取 24 小时内与之最接近的数据信息对其进行填补，若没有任何 24 小时内的数据，则抛弃该信息.



(a) 第 36534 个信息块 (b) 第 13032 个信息块

图 4: 不同填充方法比较

图4中，红色曲线为多项式预测曲线，左右两段的黑色标记点表明时间超过了 24 小时，并且可以看出多项式填补的数据严重偏离非空值，这种问题可能是由于过拟合导致，我们尝试加入正则项进行改进，得到如图4b所示，从该图中看出填补效果仍然不好，所以最终舍弃了该填补方法. 图5显示了经过填补操作后每个信息中包含的非空数据量有明显增加.

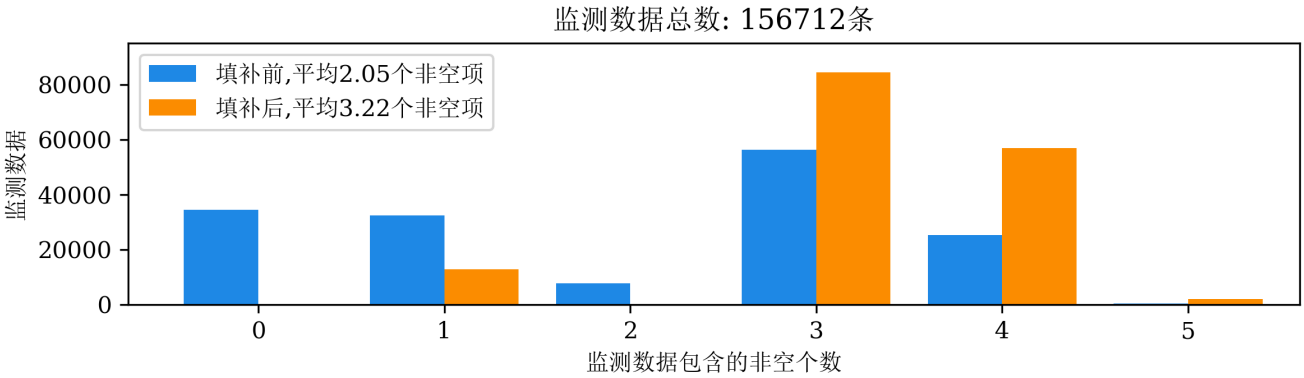


图 5: 填补前后的非空值对比

2.2.5 实时状态数据标记

根据文献 [3] 中给出的标记指标，我们可以结合 SIC 评分和 ISTH 显性 DIC 评分来对病人每个时刻是否可能患有脓毒症进行标记，具体标注规则如表4所示：

指标	评分	SIC 评分标准
PLT	2	(0, 100)
	1	[100, 150)
	0	[150, ∞)
INR	2	(1.4, ∞)
	1	(1.2, 1.4]
	0	(0, 1.2]

当 SIC 总分 ⩾ 2 时，
标记为 SIC 个体

指标	评分	DIC 评分标准
PLT	2	(0, 50)
	1	[50, 100)
	0	[100, ∞)
D-Dimer	3	[7000, ∞)
	2	[3000, 7000)
	0	(0, 3000)
FIB	1	(0,1000)
	0	[1000, ∞)
PT	2	[19, ∞)
	1	[16, 19)
	0	[0, 16)

当标记为 SIC 个体，且 DIC 总分 ⩾ 4 时，
标记为 DIC 个体

表 4: SIC 与 DIC 标注规则

使用上述填补数据进行标记后，得到的不同时间段下 SIC 和 DIC 信息占比结果由附录中表7所示.

3 模型训练与评估

3.1 数据集划分

首先需要将数据集划分为训练集与测试集，设 N_t 为序列长度，我们期望用前 N_t 个信息段的信息预测后一个信息段是否患病，设信息块为 $B = (\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \cdots, \boldsymbol{s}^{(n)})$，将 $N_t + 1$ 个时间段视为一个**时间窗口**，于是由 B 生成的时间窗口可以表示为

$$\{(\boldsymbol{s}^{(k)}, \boldsymbol{s}^{(k+1)}, \cdots, \boldsymbol{s}^{(k+N_t)}) : 1 \leqslant k \leqslant n - N_t\}$$

总计 $n - N_t$ 个时间窗口，记信息 $\boldsymbol{s}^{(k)}$ 包含 8 个维度的特征为 $\boldsymbol{s}_x^{(k)}$，将其对应的标记为 $\boldsymbol{s}_{label}^{(k)}$（如果是预测 DIC，则标记表示第 k 个信息段中是否被标记为 DIC，预测 SIC 同理），于是一个时间窗口就可以构造出如下的一组数据：

$$\begin{cases} \boldsymbol{x} = \left(\boldsymbol{s}_x^{(k)}, \boldsymbol{s}_x^{(k+1)}, \cdots, \boldsymbol{s}_x^{(k+N_t-1)}\right) \in \mathbb{R}^{8\times N_t}, \\ y = \boldsymbol{s}_{label}^{(k+N_t)}. \end{cases}$$

最后我们将数据集按照 $4:1$ **按标签比例分层抽样**，划分为训练集与测试集，数据集划分结果见附录表8.

3.2 模型评估

由于本训练集最终得到的数据量较小，所以考虑使用传统机器学习方法，我们尝试了以下 5 中算法：使用随机梯度下降优化的线性模型（SGD）、SVM 模型、决策树、随机森林、K 近邻. 并采取了以下三个评估指标对：

- 在训练集上使用 K-折叠交叉验证模型准确率.
- 在测试集上的准确率.
- ROC 曲线下面积（AUC）.

我们通过在模型在测试集上的准确率（附录表9）来对数据集的参数进行选取，对于 DIC 指标，我们发现在序列长度为 2、时间段长度为 4h 时，模型准确率较高；对于 SIC 指标，综合考虑我们发现在序列长度为 3、时间段长度为 8h 时，效果较好，其他在 98% 以上的结果我们认为是数据量较少导致的.

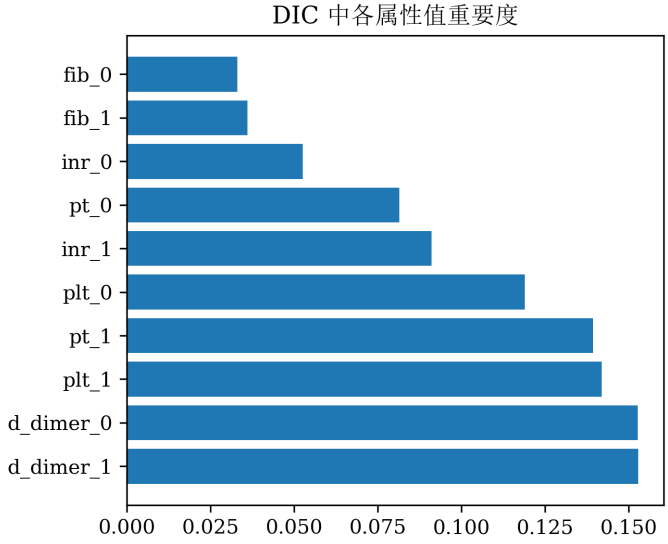
接下来我们在各自的最优数据集上使用 K-折交叉验证和 AUC 分别进一步筛选 DIC 模型和 SIC 模型，结果如表5,6所示. 并在图6中的分别展示了完整的 ROC 曲线，同时给出了基于随机森林的特征重要度分布.

模型名称	K-折交叉验证准确率	测试集准确率	ROC_AUC
SGD	[0.9151 0.8774 0.8396]	0.9125	0.9497
SVC	[0.934 0.9057 0.9245]	0.9625	0.956
决策树	[0.8774 0.9057 0.8774]	0.9875	0.8399
随机森林	[0.9245 0.9434 0.9151]	0.9875	0.9671
KNN	[0.9057 0.9245 0.8962]	0.95	0.9252

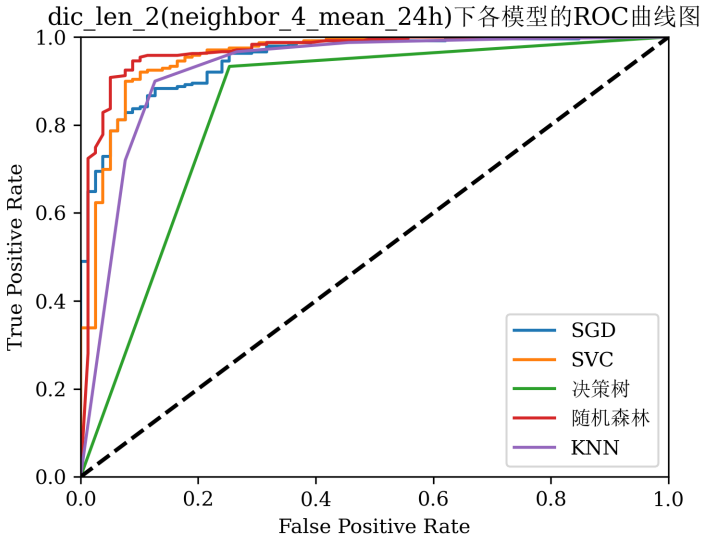
表 5: DIC 模型选择

模型名称	K-折交叉验证准确率	测试集准确率	ROC_AUC
SGD	[0.943 0.9534 0.9482]	0.9655	0.9599
SVC	[0.9326 0.9585 0.9326]	0.9586	0.9641
决策树	[0.943 0.9482 0.9223]	0.9448	0.8374
随机森林	[0.9482 0.943 0.9275]	0.9793	0.973
KNN	[0.9223 0.943 0.9482]	0.9379	0.9171

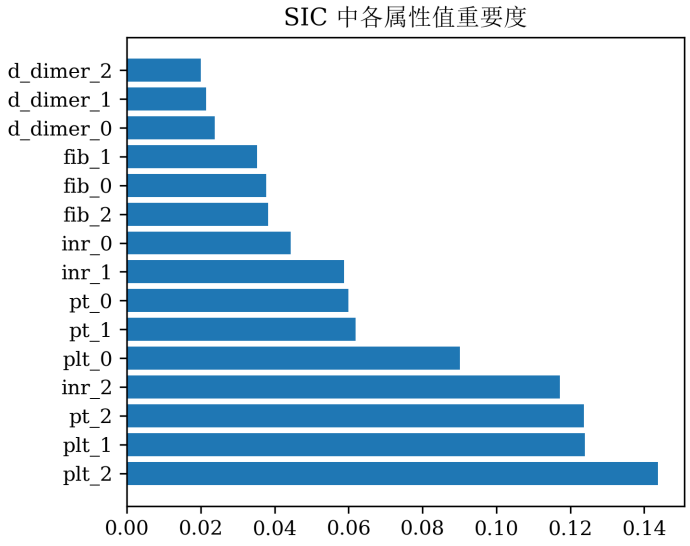
表 6: SIC 模型选择



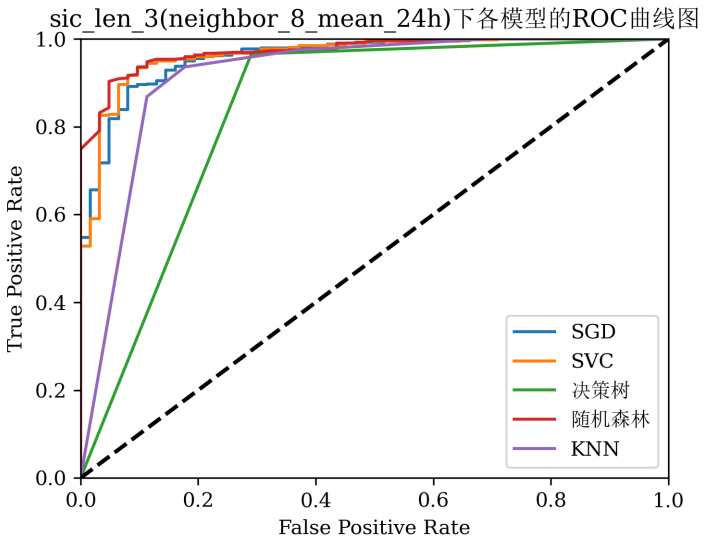
(a) DIC 重要度排名（基于随机森林）



(b) 不同模型对 DIC 预测的 ROC 值



(c) SIC 重要度排名（基于随机森林）



(d) 不同模型对 SIC 预测的 ROC 值

图 6: 不同机器学习模型分别对 DIC 和 SIC 数据集训练效果比较

4 模型可解释性分析

4.1 SHAP 值基本原理

4.1.1 加性特征归因方法

设 f 为待解释的模型，g 是用于解释该模型的模型. 解释模型通常使用简化的输入 x'，其通过函数 $h_x(x') = x$ 来映射到原本的输入. 我们主要关注的是解释单个 x 来预测 $f(x)$ 的局部方法，该方法试图保证当 $z' \approx x'$ 时，$g(z') \approx f(h_x(z'))$. 目前文献中的六种解释性方法所使用的 g 都是如下定义的加性解释模型：

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{4.1}$$

其中 $z \in \{0,1\}^M$，M 是简化输入中特征的个数，$\phi_i \in \mathbb{R}$. 称使用该类型解释模型的解释方法为加性特征归因方法.

4.1.2 加性特征归因方法中的唯一解

我们希望加性特征归因方法能够满足下列三个良好的性质，这些性质在经典的 Shapley 值估计方法中是能够保证的，但在其他方法中是未知的.

性质一：局部精度. 当对特定输入 x 近似原始模型 f 时，局部精度要求解释模型至少匹配简化输入 x'(对应于原始输入 x) 的 f 输出.

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{4.2}$$

其中 $\phi_0 = f(h_x(0))$ 表示关闭所有简化输入 (即缺失) 时的模型输出.

性质二：缺失性. 如果简化的输入中某特征不存在，那么缺失性要求原始输入中缺失的特征没有影响，具体来说即：

$$x'_i = 0 \Rightarrow \phi_i = 0 \tag{4.3}$$

当今的所有方法都满足该性质.

性质三：一致性. 该性质要求如果模型发生变化，使某些简化输入的贡献增加或保持不变，而不考虑其他输入，则该输入的归因 ϕ_i 不应减少. 具体来说，若记 $f_x(z') = f(h_x(z'))$，并令 $z'\backslash i$ 表示在 z' 中令 $z'_i = 0$，对于任意的两个模型 f 以及 f'，如果：

$$f'_x(z') - f'_x(z'\backslash i) \geqslant f_x(z') - f_x(z'\backslash i), \quad \forall z' \in \{0,1\}^M \tag{4.4}$$

那么 $\phi_i(f', x) > \phi_i(f, x)$.

以下定理表明，只有一个解释模型能够同时满足上述的三条性质:

定理 4.1. 只有一个解释模型 g 能够同时满足性质4.2,4.3以及4.4，其各个归因如下给出：

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z'\backslash i)] \tag{4.5}$$

其中 $|z'|$ 是 z' 中非零特征的个数，$z' \subseteq x'$ 表示取 x' 中非零项的子集.

定理4.1是由博弈论中的结果得出的，其中 ϕ_i 的值被称为 Shapley **值**. Young(1985) 证明 Shapley 值是满足三个性质的唯一值集.

在性质4.2-4.4下，对于给定的简化输入映射 h_x，定理4.2表明只有一种可能的加性特征归因方法. 这个结果意味着不基于 Shapley 值的方法违反了局部精度和/或一致性 (所有已知的方法都满足缺失性). 下面的章节提出了一种统一的方法来改进之前的方法，防止它们无意中违反性质4.2和4.4.

4.1.3 SHAP(SHapley Additive exPlanation) 值

在这个部分中，我们将提出 SHAP 值作为衡量特征重要度的统一度量. SHAP 值是原始模型条件期望函数的 Shapley 值，即取 $f_x(z') = f(h_x(z')) = E[f(z)|z_S]$ 时对应的 Shapley 值，其中 S 是 z' 中非零指标的集合. 由于大多数模型不能处理任意的含有缺失值的模式，我们用 $E[f(z)|z_S]$ 近似 $f(z_S)$. SHAP 值的定义旨在与 Shapley 回归、Shapley 采样紧密结合，同时还允许与 LIME、DeepLIFT 等其他加性特征归因方法相结合.

SHAP 值也可以理解成某个特征在以该特征为条件的期望模型上的变化量. 他们解释了如果我们不知道当前输出 $f(x)$ 的任何特征，如何从预测的基值 $E[f(z)]$ 得到当前输出 $f(x)$. 下图显示了特征逐步加入来计算 SHAP 值的一种排序. 然而，当模型是非线性的或输入特征不是独立的时，将特征添加到期望中的顺序很重要，并且 SHAP 值来自于对所有可能顺序的 ϕ_i 值进行平均.

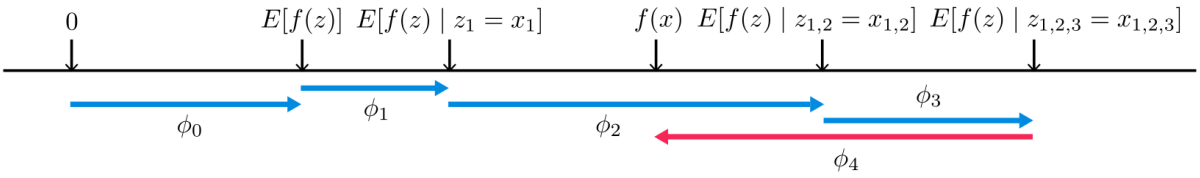


图 7: SHAP 值是由模型中的每一个特征变化得到的

SHAP 值的精确计算具有挑战性. 然而，通过结合当前加性特征归因方法的见解，我们可以近似它们. 有两种模型不可知的近似方法，一种是已知的 (Shapley 采样值)，另一种是新颖的 (Kernel SHAP). 还有四种模型类型特定的近似方法，其中两种是新颖的 (Max SHAP, Deep SHAP). 在使用这些方法时，特征之间的独立性和模型的线性性是简化期望值计算的两个可选假设 (注意 $\bar{S}$ 是不在 S 中的特征集):

$$
\begin{aligned}
f(h_x(z')) &= \mathbb{E}[f(z)|z_S] && \text{SHAP 模型简化输入映射} \\
&= \mathbb{E}_{z_{\bar{S}}|z_S}[f(z)] && \text{转化为} z_{\bar{S}}|z_S \\
&\approx \mathbb{E}_{z_{\bar{S}}}[f(z)] && \text{假设特征相互独立} \\
&\approx f(E_{z_{(\bar{S})}}[z]) && \text{假设模型线性}
\end{aligned} \tag{4.6}
$$

下面我们着重介绍模型未知时的近似计算方法.

4.1.4 LIME 解释方法

在这里着重介绍一下 LIME 解释方法，LIME 使用的局部线性解释模型完全符合方程4.1，从而是一种加性特征归因方法. LIME 将简化的输入 x' 称为“可解释输入”，映射 $x = h_x(x')$ 将可解释输入的二进制向量转换为原始输入空间. 不同类型的 h_x 映射用于不同的输入空间.

为了找到合适的 ϕ，LIME 最小化如下的目标函数:

$$\xi = \arg\min_{g \in G} L(f, g, \pi_{x'}) + \Omega(g) \tag{4.7}$$

解释模型 $g(z')$ 对原始模型 $f(h_x(z'))$ 的忠实性是通过局部核 $\pi_{x'}$ 加权的简化输入空间中一组样本上的损失 L 来实现的. Ω 惩罚了 g 的复杂性. 由于在 LIME 中，g 遵循方程4.1，L 是平方损失，因此可以使用惩罚线性回归来求解方程4.7.

4.1.5 Kernel SHAP 方法

Kernel SHAP 方法是线性 LIME 方法与 Shapley 值方法的综合. 尽管管方程4.7中 LIME 的回归公式似乎与经典 Shapley 值公式有很大的不同. 然而，由于线性 LIME 是一种加性特征归因方法，我们知道 Shapley 值是方程4.7满足三条性质（局部精度，缺失性，一致性）的唯一可能解. 一个自然的问题是，方程4.7的解是否能够恢复这些性质. 答案取决于损失函数 L、加权核 $\pi_{x'}$ 和正则化项 Ω 的选取. 下面我们将展示如何选取4.7中的参数来达到恢复 Shapley 值的效果.

定理 4.2. *能够使方程4.7的解满足性质4.2-4.4的损失函数 L, 加权核 $\pi_{x'}$ 和正则化项 Ω 为:*

$$\begin{aligned} \Omega(g) &= 0, \\ \pi_{x'}(z') &= \frac{M-1}{|z'|(M-|z'|)}, \\ L(f, g, \pi_{x'}) &= \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z'), \end{aligned}$$

注意到，当 $|z'| \in \{0, M\}$ 时，$\pi_{x'}(z') = +\infty$，从而强制地让 $\phi_0 = f_x(\varnothing)$ 和 $f(x) = \sum_{i=0}^{M} \phi_i$. 在实践中，通过使用约束条件消除这两个变量，可以在优化过程中避免这些无限权值.

由于 $g(z')$ 遵循线性形式，而 L 是平方损失，因此仍然可以使用线性回归求解方程4.7. 因此，博弈论中的 Shapley 值可以使用加权线性回归来计算. 由于 LIME 使用简化的输入映射，它相当于公式4.6中给出的 SHAP 映射的近似值，因此可以对 SHAP 值进行基于回归的、与模型无关的估计. 使用回归联合估计所有 SHAP 值比直接使用经典 Shapley 方程提供更好的样本效率（即可以用更少的样本达到比较好的精度）.

4.2 基于 SHAP 值的可解释性分析

借助 SHAP 分析工具[1]，我们对数据集分别进行了整体与单个数据的分析.

4.2.1 整体数据可解释性分析

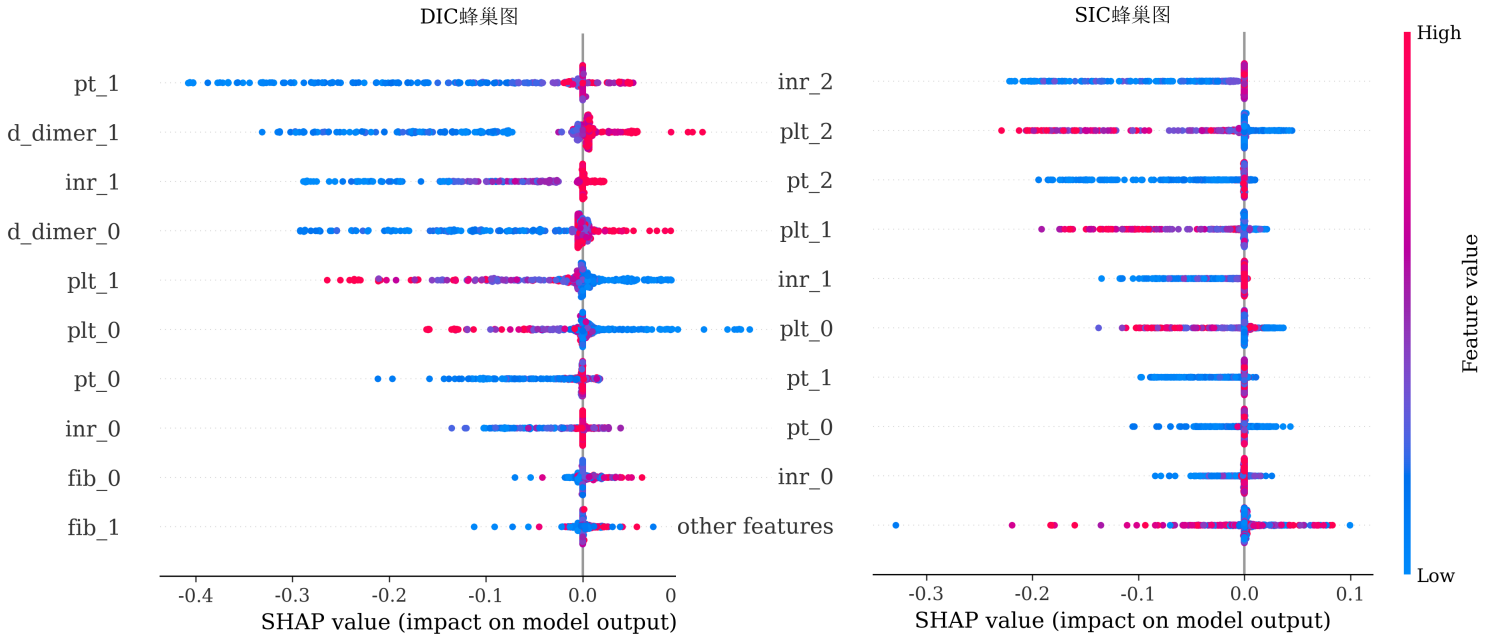
我们分别对 DIC、SIC 的两个最优数据集进行了 SHAP 值分析，整体分析结果如图8所示，这种数据分析图被形象地称作蜂巢图，其以密度的形式显示数据中的特征如何影响模型的输出，每一行表示一种特征（我们将序列长度编号从 0 开始，也就是 `pt_1` 表示第 2 个信息段中的的凝血原酶时间，如果时间从 0 开始计算，也就是 4 到 8 小时中

[1] https://github.com/slundberg/shap

数据），其上的每个点表示一个数据的该种特征对其输出的影响大小，横轴表示该特征的 SHAP 值，若某个特征在同一个 SHAP 值处存在多个数据，则会在纵轴上进行堆叠，以显示密度大小.

通过 SHAP 值的正负可以体现该特征对患病是促进还是抑制，若 SHAP 为正则说明对患病有正向作用，反之亦然. 每个数据点的颜色分布大小由右侧特征值分布柱给出，通过结合颜色与横轴，可以说明该特征的大小与患病之间的关系.

例如由图8a中第 2 行可以看出，第 2 个信息段中的 D-二聚体 `d_dim` 的与第 3 个时间段是否患 DIC 成正相关；从第 5 行可以看出，第 2 个信息段中的血小板计数 `plt` 与是否患 DIC 成负相关.



(a) DIC 特征与 SHAP 值的关系 (b) SIC 特征与 SHAP 值的关系

图 8: 两种指标下的特征与 SHAP 值之间的关系

更多地，我们通过每个特征的 SHAP 值均值大小从而用柱状图给出每个特征的重要度排名（附录图10），将其与随机森林的重要度排名图6相比，可以发现二者排名基本相同，说明 SHAP 值对模型的解释性是可性信的.

同时我们还给出了两种特征下重要度最高的两个特征的 SHAP 值与其对应取值的散点图，附录图11.

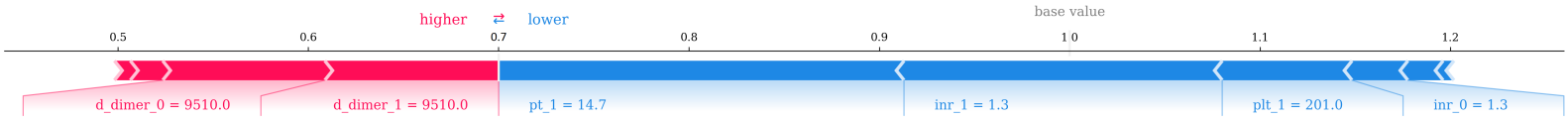
4.2.2 单个数据可解释性分析

进一步我们可以通过 SHAP 对单个病人的数据进行分析，从而说明具体哪些特征对患病具有正或负向作用，并说明当前对致病因素的影响最大的特征是什么，从而为病症的治疗给出相应的建议，图9中分别利用强度图 [5] 给出了两种标签下的影响效应（特征的定义与整体数据分析中给出的定义相同）.

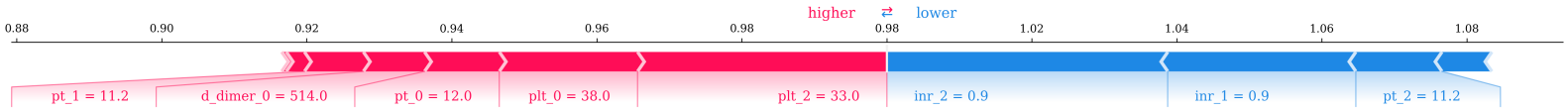
以图9a为例，该图说明模型对该病人在第 3 个时间段下患病的预测概率大小为 70%，通过 SHAP 值给出了每个特征值对预测概率的影响关系，从而我们可以根据 SHAP 值对病情给出相应的建议. 从该图中可知，第 2 个时间段下的 `d_dimer` 对患病的正促进作

用最大，而第 2 个时间段下的 `pt` 对患病的抑制作用最大，所以此时的建议应该将尝试降低 `d_dimer` 与 `pt` 为主要方向.

以下两个数据更具体影响效果请见附录中瀑布图12，瀑布图中给出了每个特征的 SHAP 值对最终预测结果的具体值大小.



(a) DIC 数据中第 104 号数据强度图分析



(b) SIC 数据中第 136 号数据强度图分析

图 9: 两种指标下的单个数据预测结果与特征之间的关系

5 结论与展望

我们以脓毒症（可诱导 SCI 、DIC ）为例，基于 MIMIC-IV 数据集，从中提取出 Platelet Count（PLT，血小板计数）、PT（凝血酶原时间）、INR(PT)（凝血酶原时间的国际标准化比值）、D-Dimer（D-二聚体）、Fibrinogen, Functional（FIB，纤维蛋白原）、pCO2（二氧化碳分压）、pO2（氧分压）和 pH（酸碱度）等与 SIC 、DIC 诊断相关的指标进行研究. 在数据预处理过程中，先利用覆盖插补和均值插补获得每个病人全部信息块，然后通过可信范围对特征的异常值进行排查，最后结合 SIC 评分和 ISTH 显性 DIC 评分来对病人每个时刻是否可能患有 SIC、DIC 进行标定，最终标记结果为（以时间段为 8h 为例）：SIC 标签数量：1885（占比 87.55%）、DIC 标签数量：1345（占比 62.47%）.

我们分别以时间段分别为 4h，8h，12h 对原数据集进行前向插补和临近填充，用前 i 个时间段的数据来预测后一个时间段的数据，这样得到的数据集分别记为 $A4_i, A8_i, A12_i$, $i = 1, 2, 3, 4$（i 为时段序列长度）. 由于数据量较小，我们使用了 5 种传统机器学习算法，分别为随机梯度下降优化的线性模型（SGD）、SVM 模型（SVC）、决策树、随机森林和 K 近邻（KNN）等，并使用了如下三个评估指标：在训练集上使用 K-折叠交叉验证准确率、测试集上的准确率、ROC 曲线下面积（AUC）. 对于 DIC 指标，我们发现在序列长度为 2、时间段长度为 4h 时，模型准确率较高；对于 SIC 指标，综合考虑我们发现在序列长度为 3、时间段长度为 8h 时，效果较好，其他在 98% 以上的结果我们认为是数据量较少导致的. 最后分别在 $A4_2$ 和 $A8_3$ 上按照先前给定的三个评估指标进一步筛选 DIC 模型和 SIC 模型，发现随机森林模型效果最好，通过随机森林同时还得出了各个属性的重要度. 最后我们使用一种极具鲁棒性的可解释方法 SHAP 来对模型的预测值进行解释，并得到了与标签值标定中相似的结果，从而可以间接地说明 SHAP 的可行性.

医疗重症监护数据稀疏性强、不规则程度高，我们通过数据插补等预处理方法提高数据使用率. 构建出可解释的实时脓毒症动态预警监测系统，以早期发现脓毒症，提高脓毒症预警模型的效能、可信度和可解释性，促进模型进入临床常规工作流程，使患者受益. 当医生之间或医生与模型决策结果不一致时，模型提供可解释的诊断预测依据显得尤为重要，可以提高模型可信度，侧面帮助医生进行诊断.

本次项目的最大不足之处就是数据集中并不自带有 DIC 和 SIC 的标签，而是通过人工打分方法标定，所以无法运用到实际医疗中，并且数据量经过筛选后数量较少，无法使用大模型进行训练；但是本次项目大幅提升了我们的代码能力，总结了很多的实验方法与思路，这些都将对我们之后做各种科研项目起到巨大的帮助作用.

参考文献

[1] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2)[DB]. PhysioNet. https://doi.org/10.13026/6mm1-ek67.

[2] Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset[DB]. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x.

[3] Toshiaki Iba, Yutaka Umemura, et al. Diagnosis of sepsis-induced disseminated intravascular coagulation and coagulopathy[J]. Acute Medicine & Surgery(July 2019), Japanese Association for Acute Medicine, Pages 223-232. https://doi.org/10.1002/ams2.411

[4] Lundberg, Scott M and Lee, Su-In. A Unified Approach to Interpreting Model Predictions[J]. Advances in Neural Information Processing Systems 30(2017), Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

[5] Scott M. Lundberg, Bala Nair, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery[J]. Nature Biomedical Engineering volume 2, pages 749 760 (2018). https://doi.org/10.1038/s41551-018-0304-0

A 附录

A.1 预处理后 SIC 和 DIC 数据量

时间段长度	插补策略	填补策略	SIC 标签数量	DIC 标签数量
4h	均值插补	临近填补	827（占比 89.89%）	629（占比 68.37%）
8h	均值插补	临近填补	1885（占比 87.55%）	1345（占比 62.47%）
12h	均值插补	临近填补	1823（占比 86.77%）	1267（占比 60.30%）

表 7: 预处理后的 SIC 和 DIC 数据量

A.2 数据集划分结果

数据集时段长度	序列长度	训练集	测试集	总计	SIC 占比	DIC 占比
4 小时	1	527	132	659	90.44%	70.41%
4 小时	2	318	80	398	92.71%	75.13%
4 小时	3	208	53	261	93.87%	78.93%
4 小时	4	132	33	165	95.15%	82.42%
8 小时	1	1308	328	1636	88.02%	63.75%
8 小时	2	895	224	1119	88.74%	65.95%
8 小时	3	579	145	724	89.36%	68.23%
8 小时	4	330	83	413	90.07%	69.73%
12 小时	1	1223	306	1529	87.25%	61.81%
12 小时	2	765	192	957	88.09%	63.64%
12 小时	3	387	97	484	89.46%	67.98%
12 小时	4	202	51	253	90.51%	70.75%

表 8: 数据集划分结果（均使用临近填充与均值插补）

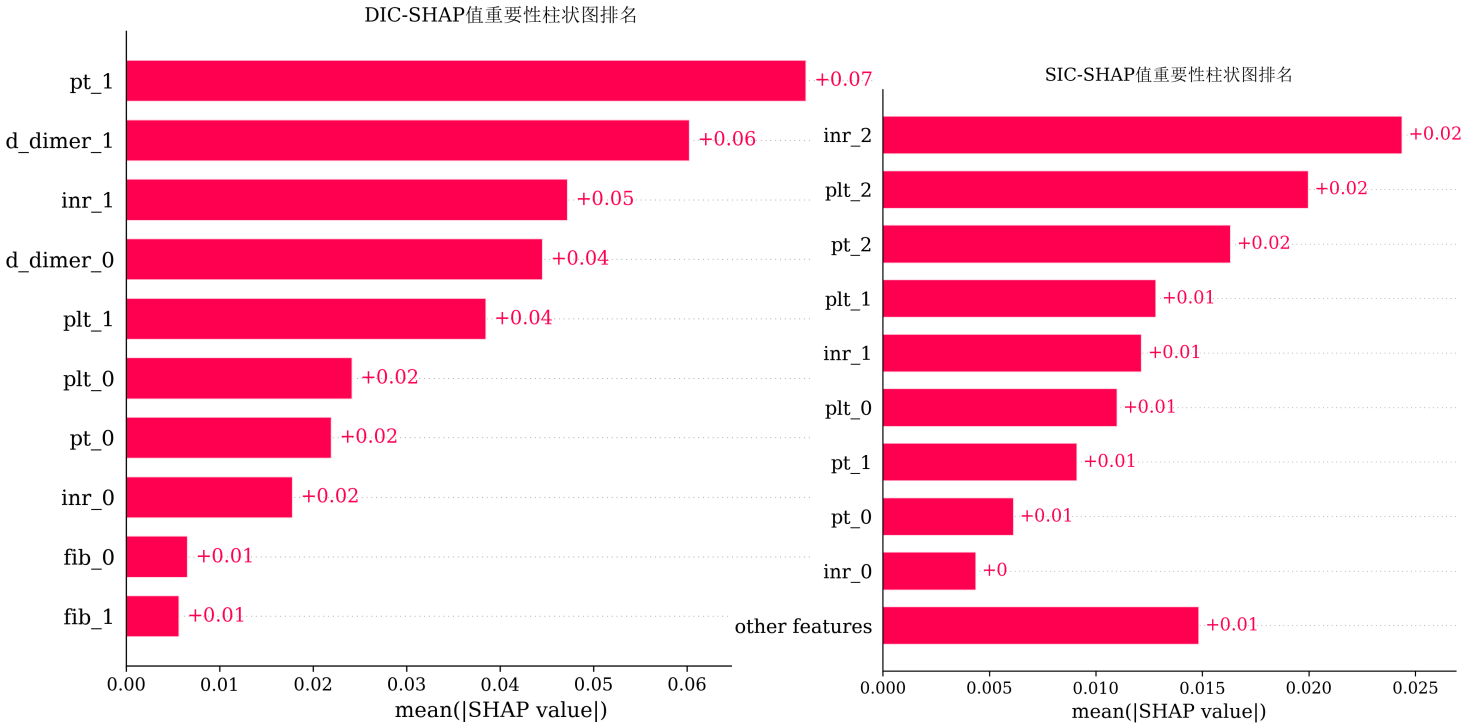
A.3 模型评估结果

数据集名称解释：“dic_len_1(neighbor_12_mean_24h)”表示预测指标为 DIC，基于前 1 个 12 小时的信息段，预测后 12 小时是否患有 DIC，使用临近填充和均值插补.

数据集名称	SGD	SVC	决策树	随机森林	KNN
dic_len_1(neighbor_12_mean_24h)	0.8497	0.8922	0.8856	0.8954	0.8693
dic_len_1(neighbor_4_mean_24h)	0.8712	0.9394	0.9318	0.9394	0.8788
dic_len_1(neighbor_8_mean_24h)	0.878	0.9482	0.939	0.9482	0.9116
dic_len_2(neighbor_12_mean_24h)	0.901	0.9115	0.901	0.9323	0.9115
dic_len_2(neighbor_4_mean_24h)	0.9125	0.9625	0.9875	0.9875	0.95
dic_len_2(neighbor_8_mean_24h)	0.8616	0.9286	0.9196	0.9286	0.933
dic_len_3(neighbor_12_mean_24h)	0.8557	0.9278	0.8763	0.9588	0.8969
dic_len_3(neighbor_4_mean_24h)	0.8868	0.9434	0.9434	0.9434	0.9245
dic_len_3(neighbor_8_mean_24h)	0.9034	0.9448	0.9448	0.9793	0.9448
dic_len_4(neighbor_12_mean_24h)	0.8431	0.9412	0.9608	0.9804	0.9216
dic_len_4(neighbor_4_mean_24h)	0.9394	0.9091	0.8182	0.9394	0.9091
dic_len_4(neighbor_8_mean_24h)	0.9157	0.9398	0.9759	0.9759	0.9277
sic_len_1(neighbor_12_mean_24h)	0.8987	0.9248	0.9183	0.9346	0.9346
sic_len_1(neighbor_4_mean_24h)	0.947	0.9621	0.9394	0.9545	0.9394
sic_len_1(neighbor_8_mean_24h)	0.9329	0.9634	0.9543	0.9665	0.9573
sic_len_2(neighbor_12_mean_24h)	0.8854	0.9375	0.9062	0.9427	0.9375
sic_len_2(neighbor_4_mean_24h)	0.9625	0.975	0.9625	0.975	0.95
sic_len_2(neighbor_8_mean_24h)	0.9464	0.9643	0.9554	0.9509	0.9241
sic_len_3(neighbor_12_mean_24h)	0.8969	0.9485	0.9278	0.9588	0.9485
sic_len_3(neighbor_4_mean_24h)	0.9811	0.9811	0.9811	0.9811	0.9811
sic_len_3(neighbor_8_mean_24h)	0.9655	0.9586	0.9448	0.9793	0.9379
sic_len_4(neighbor_12_mean_24h)	0.9412	0.9608	0.9804	1.0	0.9804
sic_len_4(neighbor_4_mean_24h)	0.9091	1.0	0.9697	1.0	0.9697
sic_len_4(neighbor_8_mean_24h)	0.9277	0.9398	0.9277	0.9639	0.9398

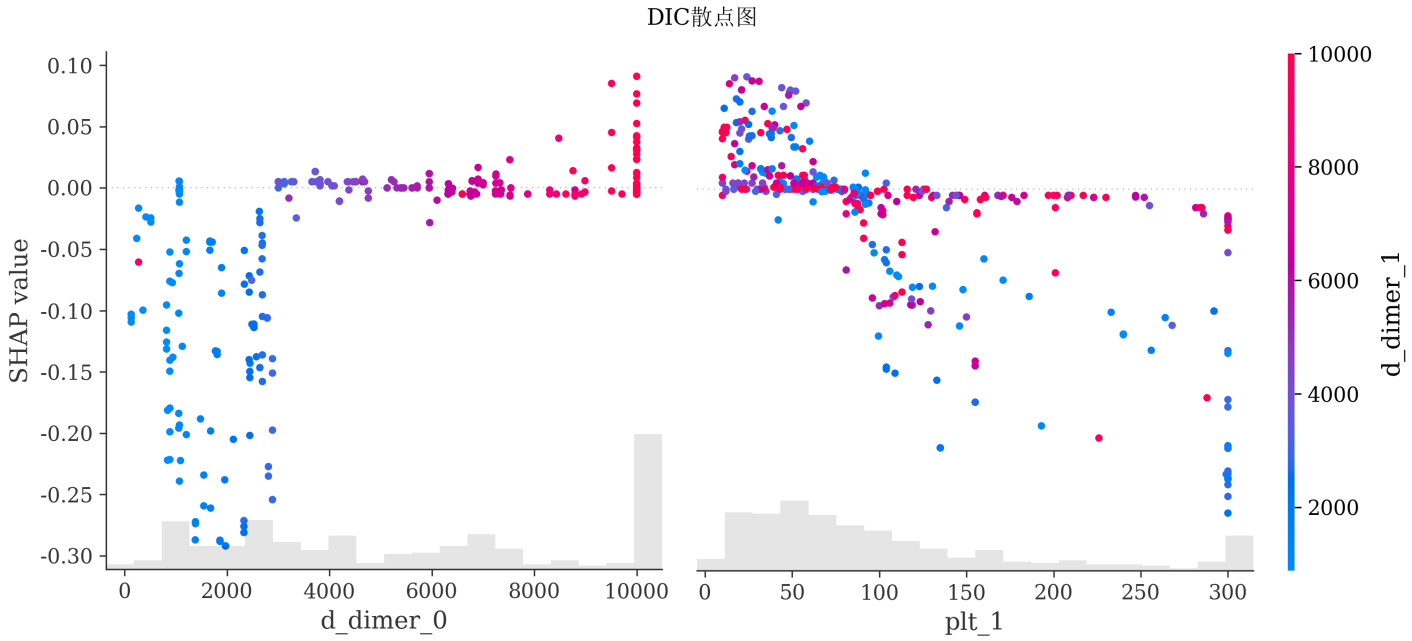
表 9: 五种机器学习模型评估结果比对

A.4 基于 SHAP 值的可解释性分析

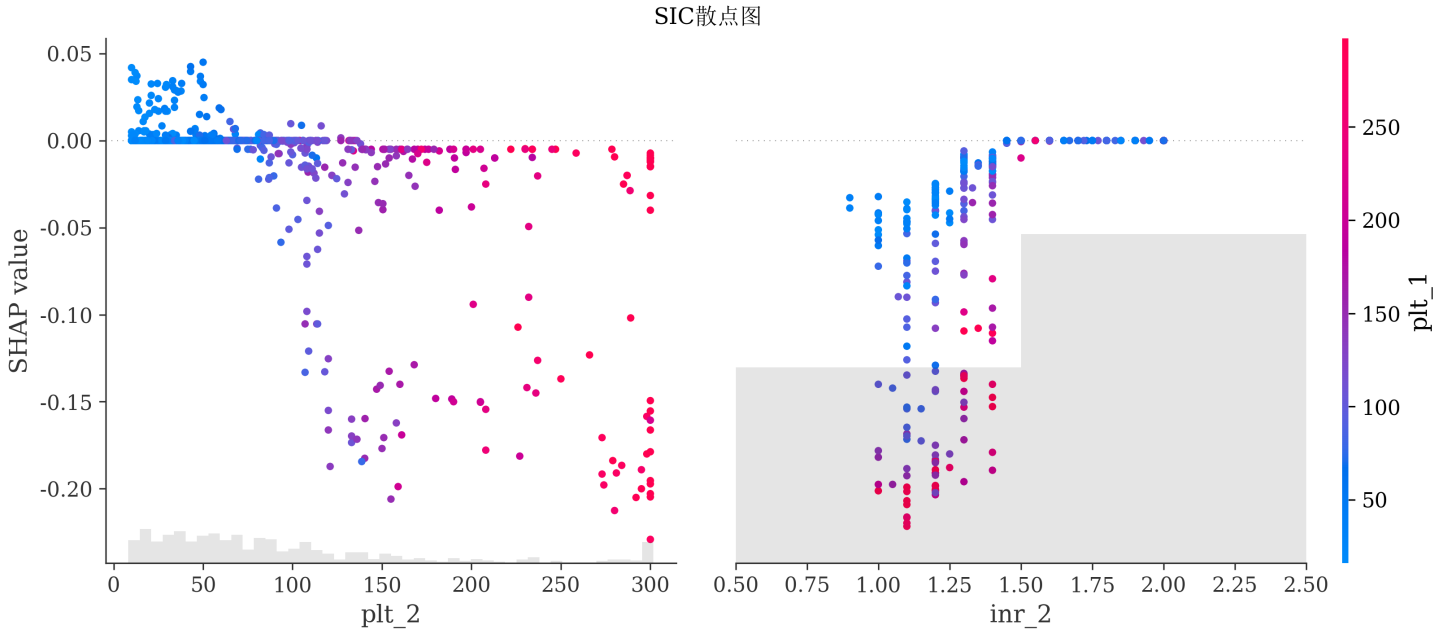


(a) DIC 特征基于 SHAP 值的重要度　　(b) SIC 特征基于 SHAP 值的重要度

图 10: 两种指标下的基于 SHAP 的特征重要度大小

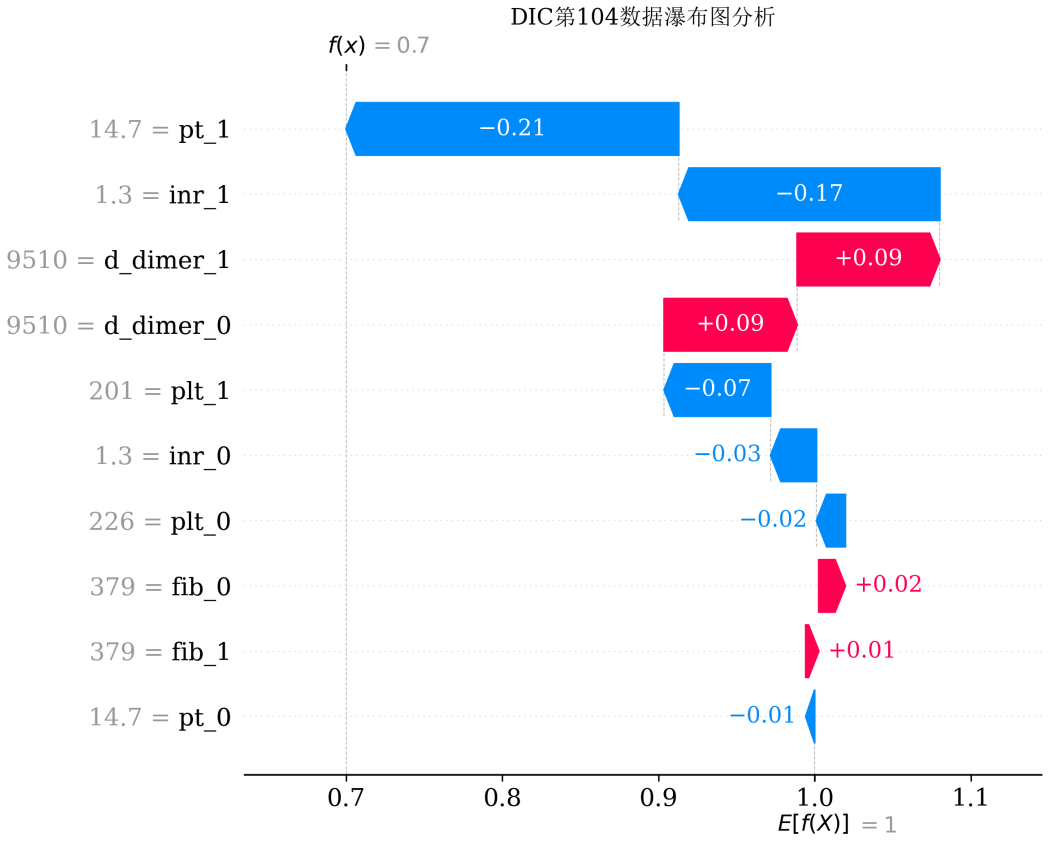


(a) DIC 中的两个关键特征与 SHAP 值的关系

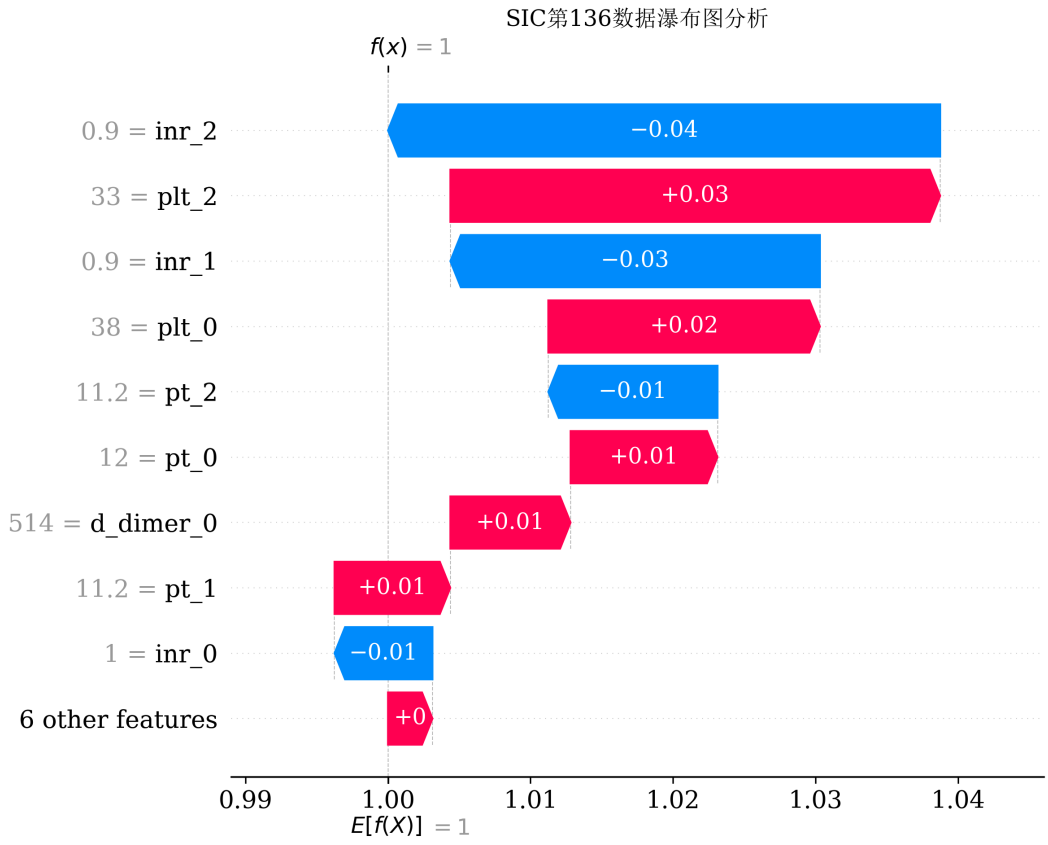


(b) SIC 中的两个关键特征与 SHAP 值的关系

图 11: 两种指标下的关键特征与 SHAP 值之间的关系



(a) DIC 数据中第 104 号数据瀑布图分析



(b) SIC 数据中第 136 号数据瀑布图分析

图 12: 两种指标下的单个数据预测结果与特征之间的关系（瀑布图）