

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：yupeng666@stu.xjtu.edu.cn

项目成员：吴天阳，王承杰

指导老师：孙剑

批准经费：10000 已用金额：631.12

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

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	1
2	数据集分析及预处理	2
2.1	数据集介绍	2
2.2	数据预处理	2
2.2.1	数据提取	2
2.2.2	数据插补	3
2.2.3	异常值矫正	5
2.2.4	数据填补	5
2.2.5	实时状态数据标记	7
3	模型训练与评估	8
3.1	数据集划分	8
3.2	模型评估	8
4	模型可解释性分析	10
4.1	SHAP 值基本原理	10
4.2	基于 SHAP 值的可解释性分析	10
4.2.1	整体数据可解释性分析	10
4.2.2	单个数据可解释性分析	11
5	结论与展望	11
A	附录	13
A.1	预处理后 SIC 和 DIC 数据量	13
A.2	数据集划分结果	13
A.3	模型评估结果	14
A.4	基于 SHAP 值的可解释性分析	15

# 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 sepsis predictor, 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 Record 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.

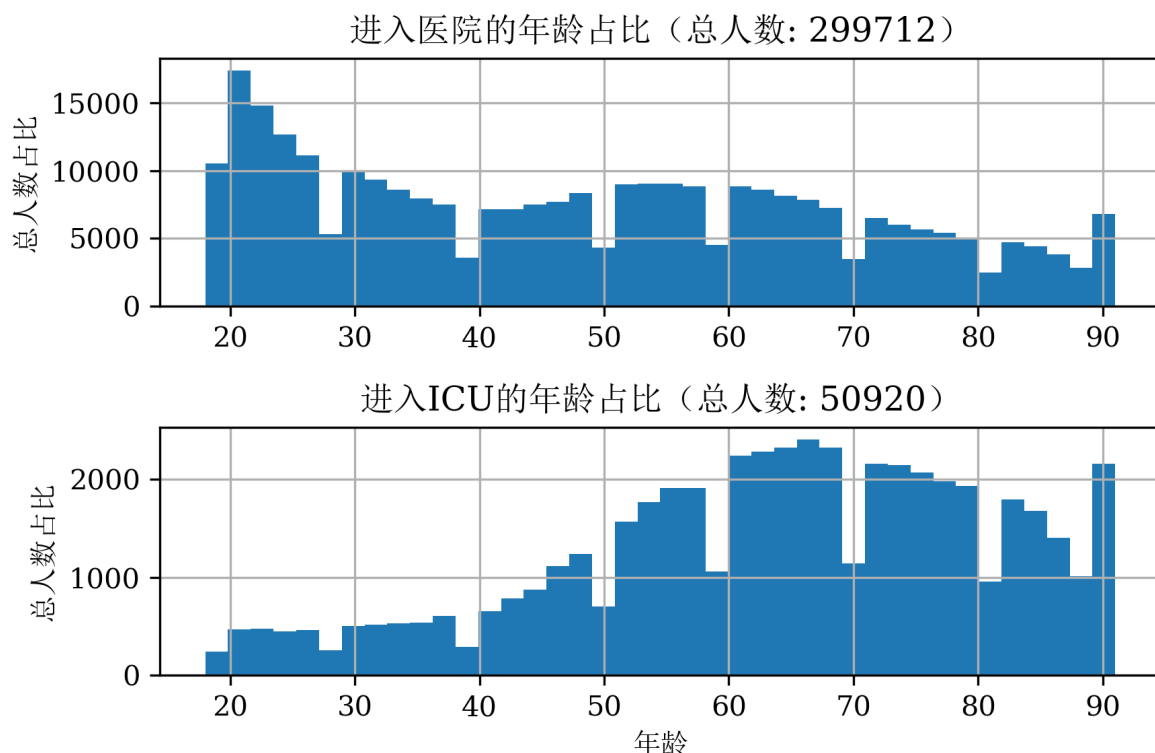


图 1: 数据集中年龄占比分布

### 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短对数据进行插补。

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

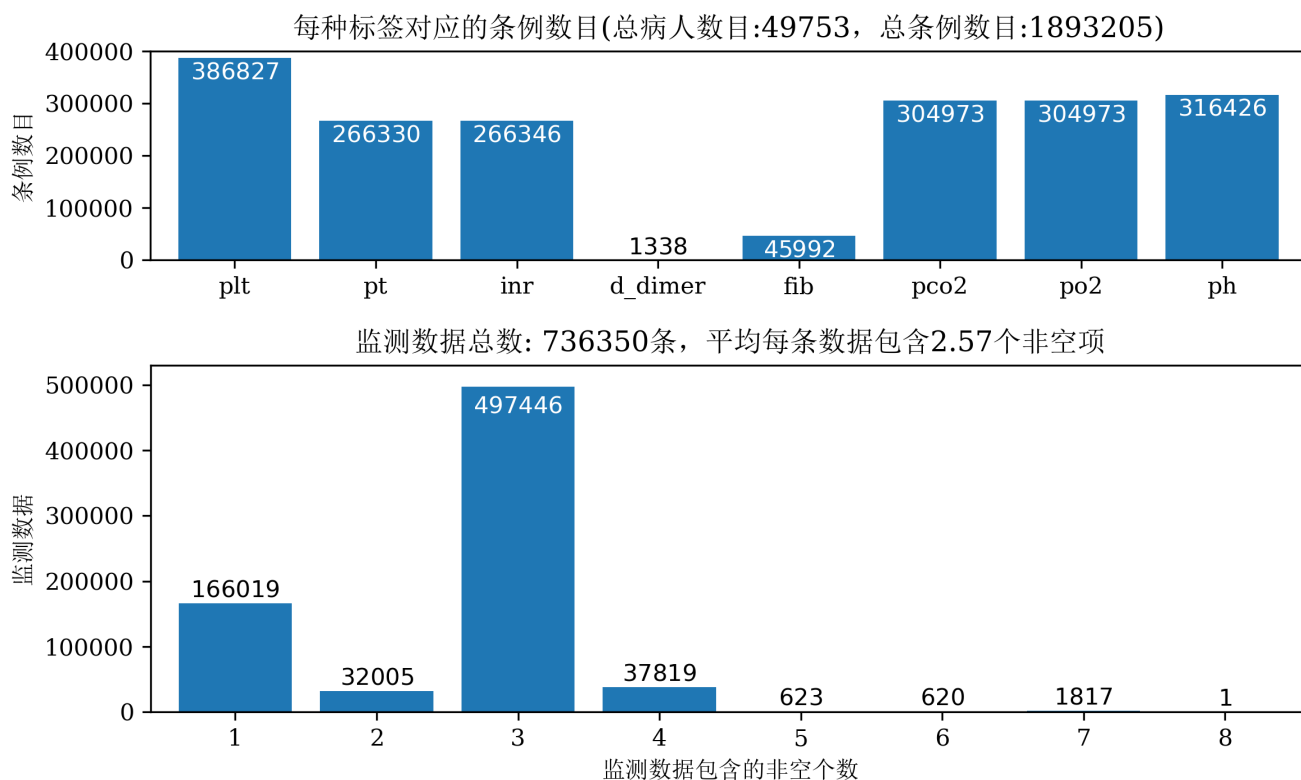


图 2: 数据提取结果

### 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

```

1 charttimes = df['charttime']
2 # 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
3 time_format = '%Y-%m-%d %H:%M:%S'
4 base_time = datetime.strptime(charttimes.min(), time_format)
5
6 # 计算与基准时间的相对时间差 (单位: 小时, 浮点型)
7 def relative_hours(date_string):
8     date = datetime.strptime(date_string, time_format)
9     relative_days = (date - base_time).days # 相对天数
10    relative_seconds = (date - base_time).seconds # 相对秒数
11    relative_hours = 24 * relative_days + relative_seconds / 3600
12    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  $i \in \{1, \dots, N\}$  条信息记为  $\mathbf{x}^{(i)} = \{x_k^{(i)}\}_{k=1}^8 \in \mathbb{R}^8$ , 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  $\mathbf{y}^{(j)} = \{y_k^{(j)}\}_{k=1}^8 \in \mathbb{R}^8$ . 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:

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

**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:

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

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 $\mathbf{x}^{(i)}$ 对应的时刻
$S = (\mathbf{x}^{(i_1)}, \dots, \mathbf{x}^{(i_k)})$	信息段: 按照信息的时间戳从小到大排序, 并且满足 $x_{time}^{(i_k)} - x_{time}^{(i_1)} \leq L$
$S_1 \prec S_2$	信息段的序关系: $\mathbf{x}_{time} < \mathbf{y}_{time}, (\mathbf{x} \in S_1, \mathbf{y} \in S_2)$
$S_1 \sim S_2$	连续信息段: $\min_{\mathbf{y} \in S_2} y_{time} - \max_{\mathbf{x} \in S_1} x_{time} \leq L, (S_1 \prec S_2)$
$S_1 \sim S_2 \sim \dots \sim S_k$	多个连续信息段: $S_i \sim S_{i+1}, (i = 1, \dots, k-1)$
$B = \{S_1, S_2, \dots, S_k\}$	信息块: $S_1 \sim \dots \sim S_k$ 且 $k \geq 2$

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表情对应的非空信息占比如图3

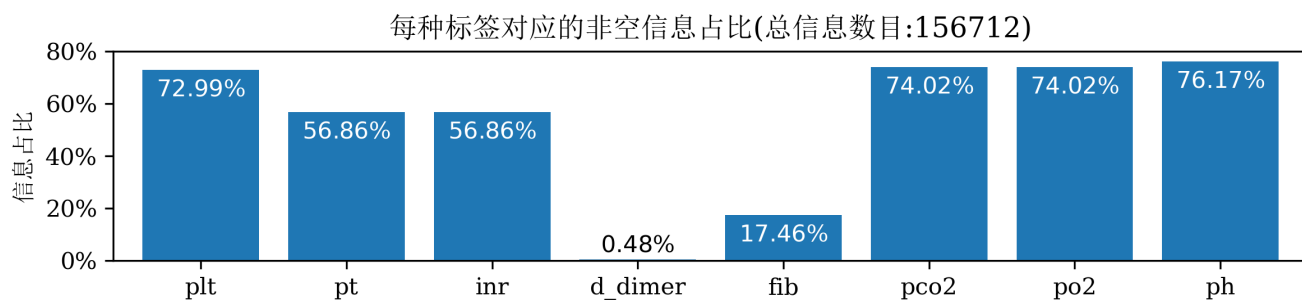


图 3: 插补操作后非空信息占比

### 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	$\leq 16$	[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。

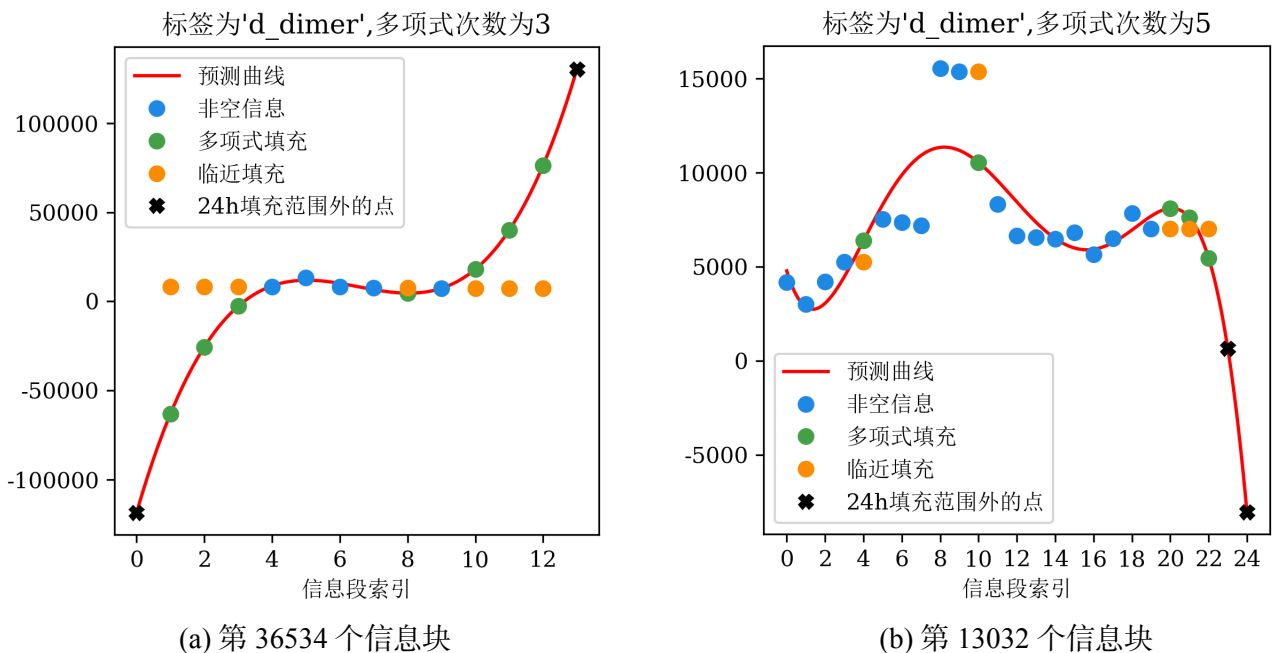


图 4: 不同填充方法比较

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.

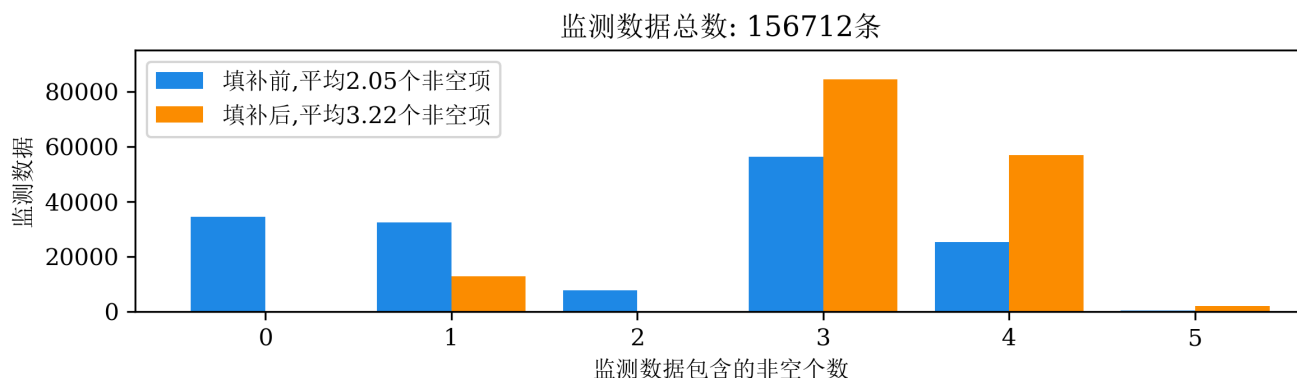


图 5: 填补前后的非空值对比

### 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	指标	评分	DIC 评分标准
PLT	2	(0, 100)	PLT	2	(0, 50)
	1	[100, 150)		1	[50, 100)
	0	[150, $\infty$ )		0	[100, $\infty$ )
INR	2	(1.4, $\infty$ )	D-Dimer	3	[7000, $\infty$ )
	1	(1.2, 1.4]		2	[3000, 7000)
	0	(0, 1.2]		0	(0, 3000)
FIB	1		FIB	1	(0,1000)
	0			0	[1000, $\infty$ )
PT	2		PT	2	[19, $\infty$ )
	1			1	[16, 19)
	0			0	[0, 16)

当 SIC 总分  $\geq 2$  时,  
标记为 SIC 个体

当标记为 SIC 个体, 且 DIC 总分  $\geq 4$  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  $B = (s^{(1)}, s^{(2)}, \dots, s^{(n)})$ ，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

$$\{(s^{(k)}, s^{(k+1)}, \dots, s^{(k+N_t)}) : 1 \leq k \leq n - N_t\}$$

总计  $n - N_t$  个时间窗口，记信息  $s^{(k)}$  包含 8 个维度的特征为  $s_x^{(k)}$ ，将其对应的标记为  $s_{label}^{(k)}$ （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：

$$\begin{cases} \mathbf{x} = (s_x^{(k)}, s_x^{(k+1)}, \dots, s_x^{(k+N_t-1)}) \in \mathbb{R}^{8 \times N_t}, \\ y = s_{label}^{(k+N_t)}. \end{cases}$$

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。

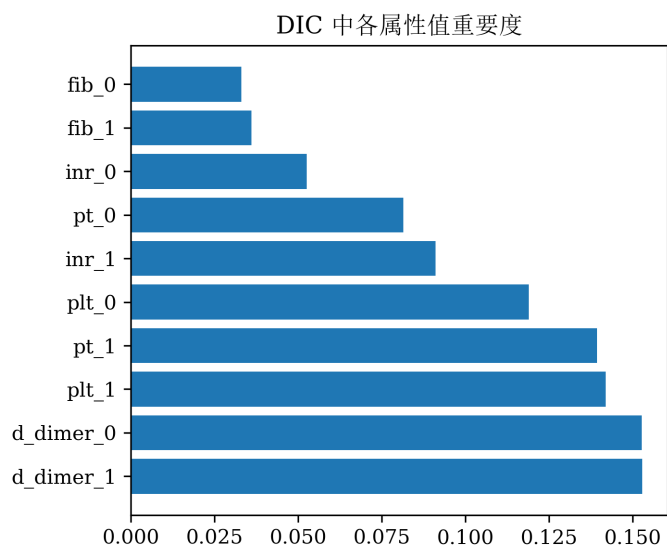
接下来我们在各自的最优数据集上使用 K-折交叉验证和 AUC 分别进一步筛选 DIC 模型和 SIC 模型，结果如表5,6所示。并在图6中的分别展示了完整的 ROC 曲线，同时给出了基于随机森林的特征重要度分布。

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

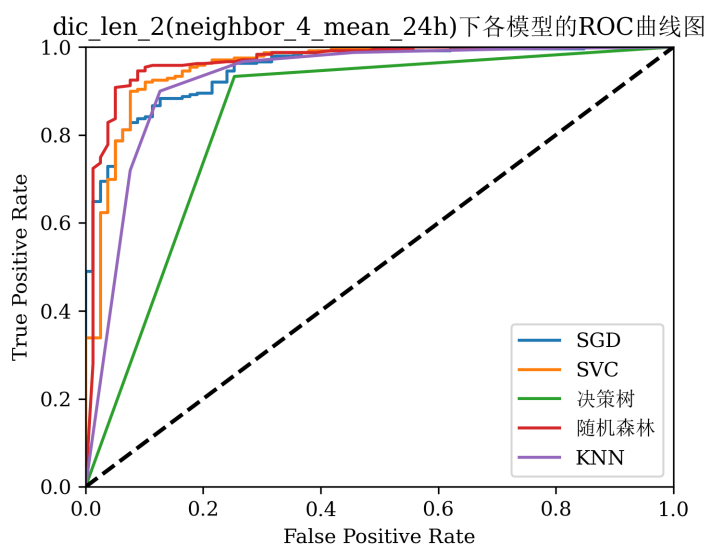
表 5: DIC 模型选择

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

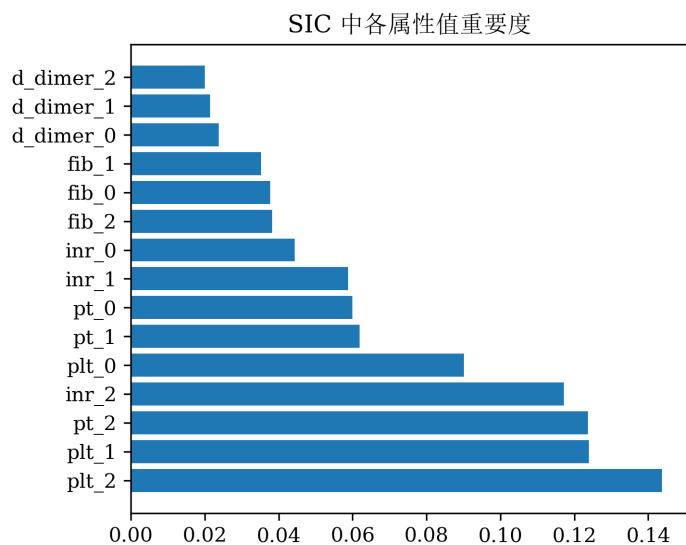
表 6: SIC 模型选择



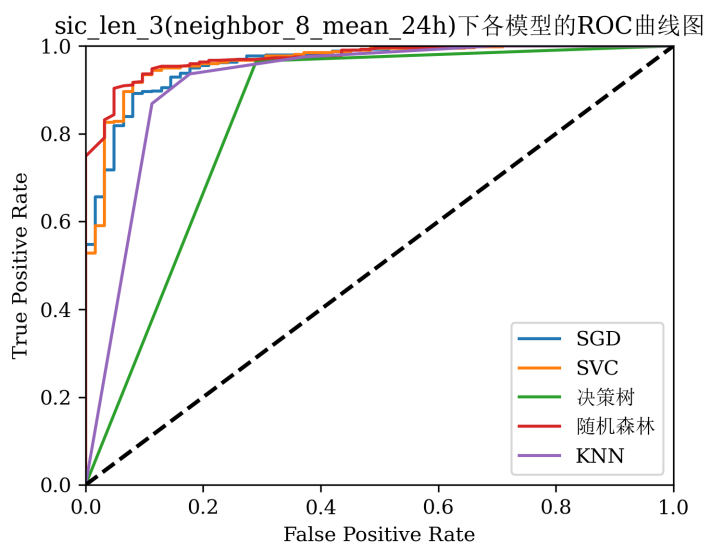
(a) DIC 重要度排名 (基于随机森林)



(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

### 4.2 基于 SHAP 值的可解释性分析

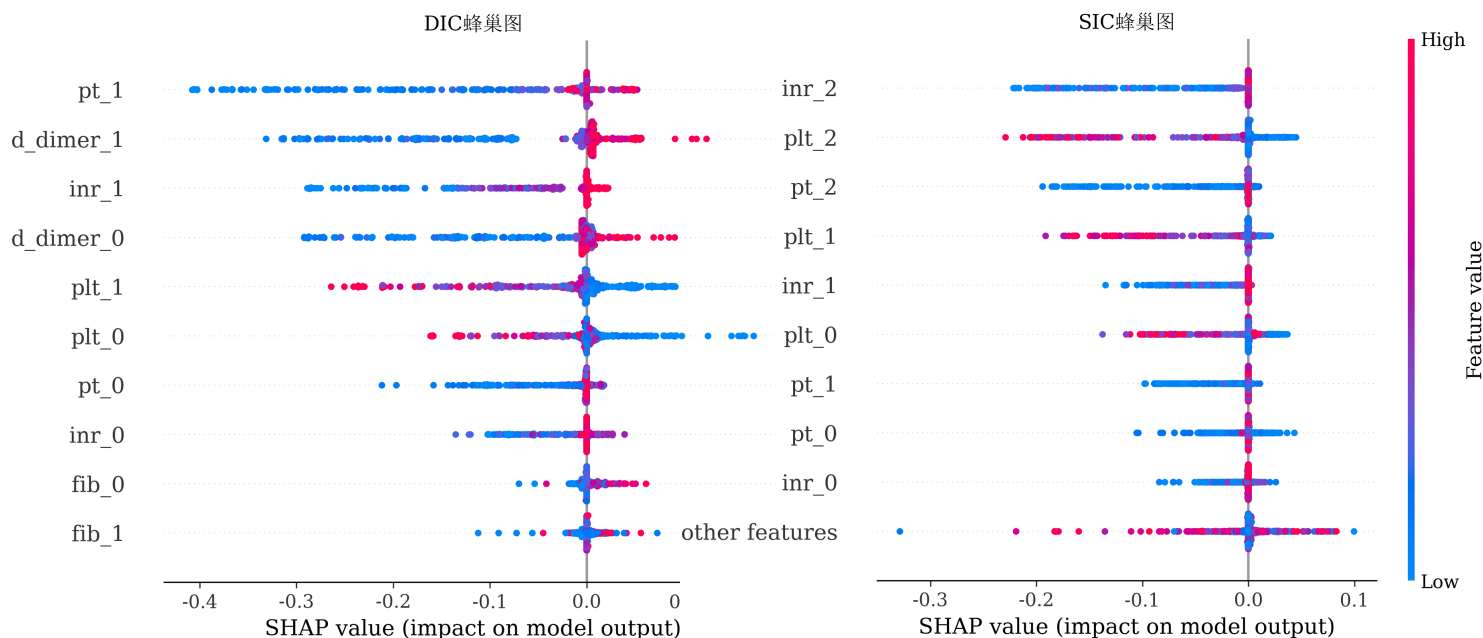
借助 SHAP 分析工具<sup>1</sup>，我们对数据集分别进行了整体与单个数据的分析。

#### 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7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7a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 7: 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

<sup>1</sup><https://github.com/slundberg/shap>

度排名（附录图9），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10。

#### 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8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8a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11，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。



图 8: 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

通过本次大创项目，我们学习到了如何高效地处理庞大的数据集，并且对数据进行清洗，利用时间序列的方法对数据进行划分和特征提取，在对异常值进行处理后，我们设计了不同的插补和缺失值填充方法对数据进行扩充；经过数据预处理之后，我们又使用了多种机器学习方法进行模型构建，并相互之间进行了比较，从而选出了最优的模型与数据集；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。

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，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)



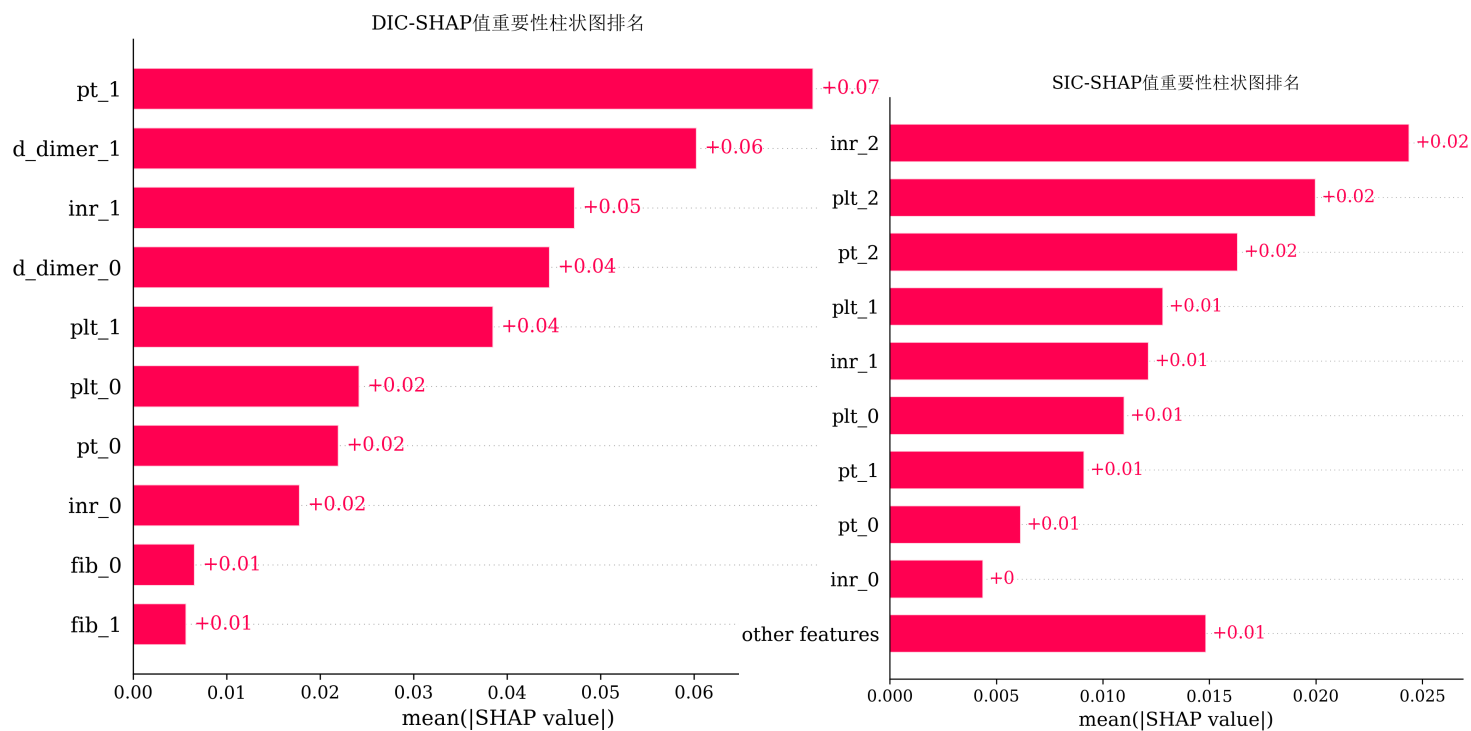
### A.3 模型评估结果

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

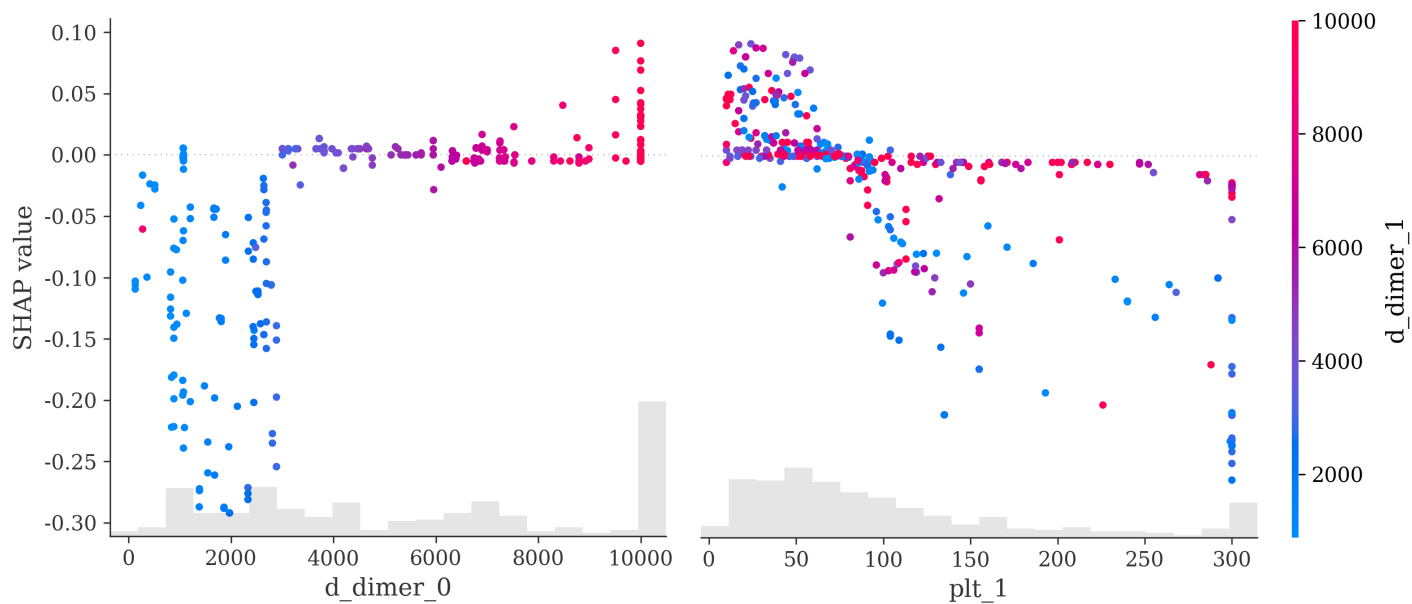


(a) DIC 特征基于 SHAP 值的重要度

(b) SIC 特征基于 SHAP 值的重要度

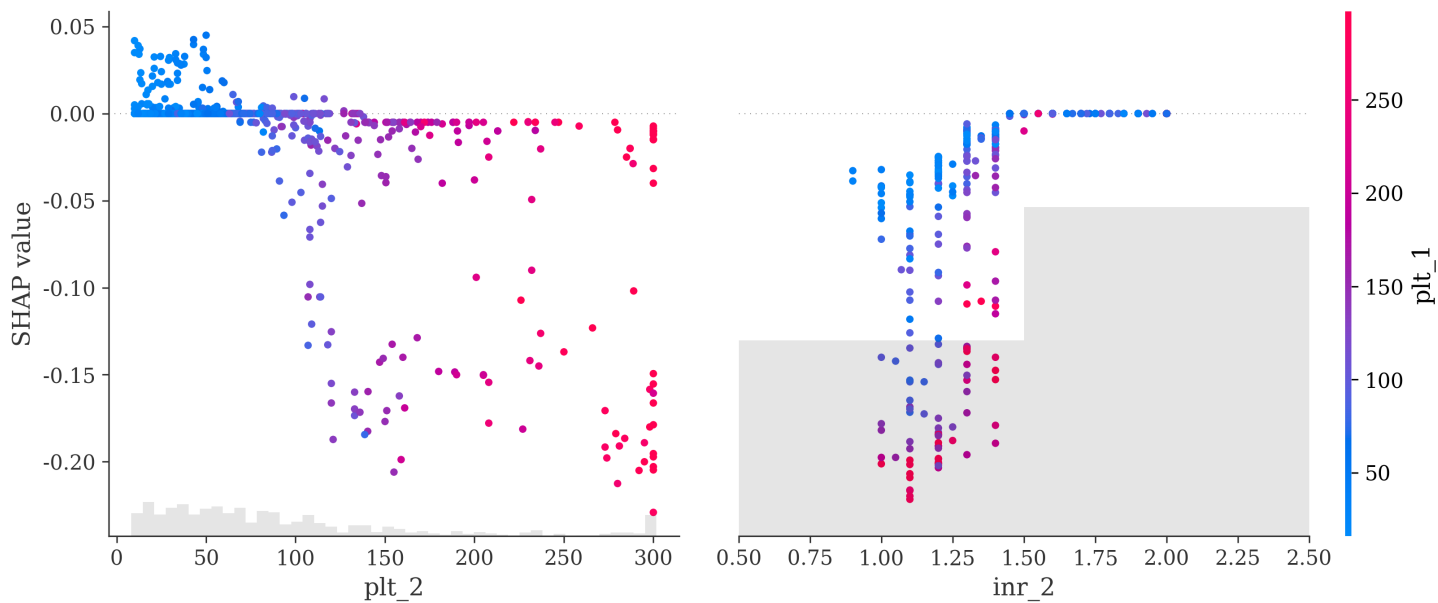
图 9: 两种指标下的基于 SHAP 的特征重要度大小

DIC散点图



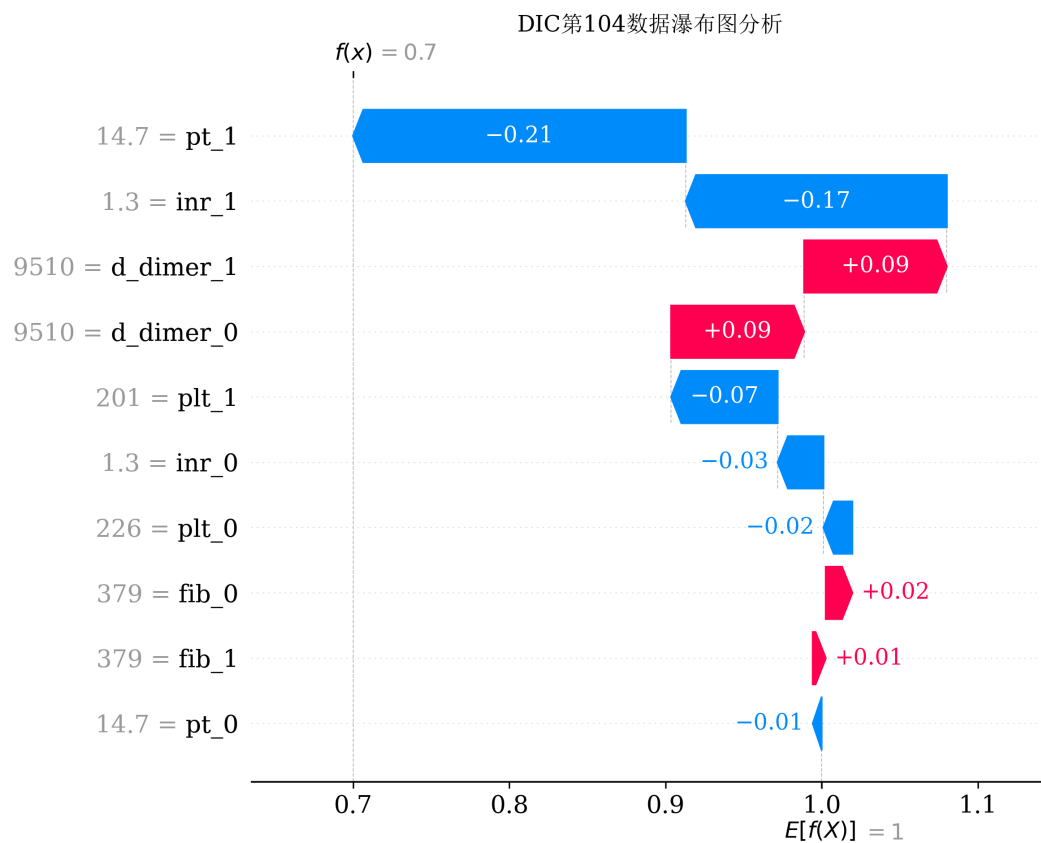
(a) DIC 中的两个关键特征与 SHAP 值的关系

SIC散点图

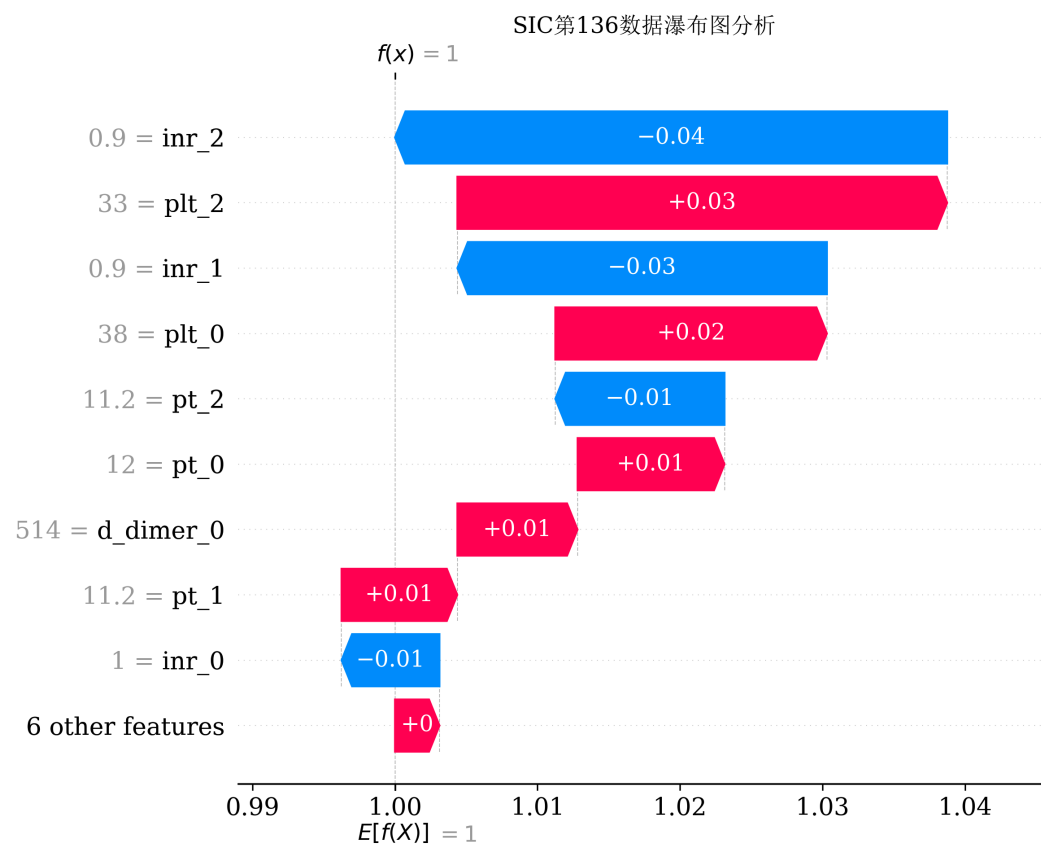


(b) SIC 中的两个关键特征与 SHAP 值的关系

图 10: 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 11: 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）